

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Volume I

1941



Published by

SCIENCE RESEARCH ASSOCIATES

1700 PRAIRIE AVENUE • CHICAGO, ILLINOIS

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

A quarterly journal devoted to the development and application of
measures of individual differences.

EDITOR

G. FREDERIC KUDER. Social Security Board

ASSOCIATE EDITORS

DOROTHY C. ADKINS. Social Security Board

FORREST A. KINGSBURY. University of Chicago

M. W. RICHARDSON. United States Civil Service Commission

BOARD OF COOPERATING EDITORS

RICHARD D. ALLEN
Providence Public Schools

P. J. RULON
Harvard University

JOHN G. DARLEY
University of Minnesota

DAVID SEGEL
U. S. Office of Education

HAROLD A. EDGERTON
Ohio State University

C. L. SHARTLE
Social Security Board

MAX D. ENGELHART
Chicago City Junior Colleges

H. C. TAYLOR
Western Electric Company

E. B. GREENE
University of Michigan

THELMA G. THURSTONE
Chicago Teachers College

J. P. GUILFORD
University of Southern California

HERBERT A. TOOPS
Ohio State University

E. F. LINDQUIST
State University of Iowa

E. G. WILLIAMSON
University of Minnesota

BEN D. WOOD
Columbia University

The journal is open to (1) reports of research on the development and use of tests and measurements in education, government, and industry, (2) descriptions of testing programs being used for various purposes, (3) discussions of problems of measurement in general or in specific fields, and (4) miscellaneous notes pertinent to the measurement field, such as suggestions of new types of items or improved methods of treating test data. Manuscripts should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1700 Prairie Avenue, Chicago, Illinois

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly by Science Research Associates, 1700 Prairie Avenue, Chicago, Illinois. Subscription rate, \$4.00 a year. Entered as second class matter June 11, 1941, at the Post Office at Chicago, Illinois, under the Act of March 3, 1879.

INDEX FOR VOLUME I

<i>Andrus, Lawrence</i> A COMPOSITION TEST FOR FOREIGN LANGUAGES.....	355
<i>Blakey, Robert I.</i> A FACTOR ANALYSIS OF A NON-VERBAL REASONING TEST. . .	187
<i>Bordin, E. S. and Williamson, E. G.</i> THE EVALUATION OF VOCATIONAL AND EDUCATIONAL COUNSEL- ING: A CRITIQUE OF THE METHODOLOGY OF EXPERIMENTS....	5
<i>Bordin, E. S. and Williamson, E. G.</i> AN ANALYTICAL DESCRIPTION OF STUDENT COUNSELING.....	341
<i>Bordin, E. S. and Sarbin, T. R.</i> NEW CRITERIA FOR OLD.	173
<i>Cronbach, Lee J.</i> THE RELIABILITY OF RATIO SCORES.....	269
<i>Darley, John G.</i> COUNSELING ON THE BASIS OF INTEREST MEASUREMENT.	35
<i>Edgerton, Harold A. and Ellison, Mary Lou</i> THE THURSTONE PRIMARY MENTAL ABILITIES TEST AND COL- LEGE MARKS	399
<i>Ellison, Mary Lou and Edgerton, Harold A.</i> THE THURSTONE PRIMARY MENTAL ABILITIES TEST AND COL- LEGE MARKS	399
<i>Engelhart, Max D. and Lewis, Hugh B.</i> AN ATTEMPT TO MEASURE SCIENTIFIC THINKING.	289
<i>Harrell, Willard and Faubion, Richard</i> PRIMARY MENTAL ABILITIES AND AVIATION MAINTENANCE COURSES	59
<i>Hartson, L. D. and Sprovo, A. J.</i> THE VALUE OF INTELLIGENCE QUOTIENTS OBTAINED IN SEC- ONDARY SCHOOL FOR PREDICTING COLLEGE SCHOLARSHIP.....	387
<i>Huskey, Marshall S.</i> A NEW PERFORMANCE TEST FOR YOUNG DEAF CHILDREN.....	217
<i>Hoyt, G. J.</i> NOTE ON A SIMPLIFIED METHOD OF COMPUTING TEST RELIA- BILITY	93
<i>Kopas, Joseph S.</i> GUIDING STUDENTS TO BECOME SELF-GUIDING.....	279
<i>Koran, Sidney W.</i> PERFORMANCE TESTING IN PUBLIC PERSONNEL SELECTION, PART I	233
<i>Koran, Sidney W.</i> PERFORMANCE TESTING IN PUBLIC PERSONNEL SELECTION, PART II	365
<i>Kuder, G. Frederic and Shanner, William M.</i> A COMPARATIVE STUDY OF FRESHMAN WEEK TESTS GIVEN AT THE UNIVERSITY OF CHICAGO.....	85
<i>Lewis, Hugh B. and Engelhart, Max D.</i> AN ATTEMPT TO MEASURE SCIENTIFIC THINKING.....	289
<i>Lorr, Maurice and Meister, Ralph K.</i> THE CONCEPT OF SCATTER IN THE LIGHT OF MENTAL TEST THEORY	303
<i>McCall, William G. and Traxler, Arthur E.</i> SOME DATA ON THE KUDER PREFERENCE RECORD	253

THE CONCEPT OF SCATTER IN THE LIGHT OF MENTAL TEST THEORY	303
<i>Meister, Ralph K. and Reymert, Martin L.</i>	
A COMPARISON OF THE ORIGINAL AND REVISED STANFORD-BINET INTELLIGENCE SCALES	67
<i>Mosier, Charles I.</i>	
A SHORT CUT IN THE ESTIMATION OF SPLIT-HALVES COEFFICIENTS	407
<i>Munson, Grace</i>	
THE COURSE IN SELF-APPRAISAL AND CAREERS OFFERED TO SENIORS IN CHICAGO PUBLIC SCHOOLS.....	43
<i>Powell, Norman J.</i>	
EXAMINING EXAMINERS	187
<i>Reymert, Martin L. and Meister, Ralph K.</i>	
A COMPARISON OF THE ORIGINAL AND REVISED STANFORD-BINET INTELLIGENT SCALES	67
<i>Richardson, M. W.</i>	
THE LOGIC OF AGE SCALES.....	25
<i>Sandt, Karl E. and Triggs, Frances Oralind</i>	
AN EVALUATION OF TECHNIQUES OF MEASURING VISUAL ACTIVITY AT THE COLLEGE LEVEL.....	295
<i>Sarbin, T. R. and Bordin, E. S.</i>	
NEW CRITERIA FOR OLD.....	173
<i>Schneidler, Gwendolen G.</i>	
GRADE AND AGE NORMS FOR THE MINNESOTA VOCATIONAL TEST FOR CLERICAL WORKERS	143
<i>Shanner, William M. and Kuder, G. Frederic</i>	
A COMPARATIVE STUDY OF FRESHMAN WEEK TESTS GIVEN AT THE UNIVERSITY OF CHICAGO.....	85
<i>Sprow, A. J. and Hartson, L. D.</i>	
THE VALUE OF INTELLIGENCE QUOTIENTS OBTAINED IN SECONDARY SCHOOL FOR PREDICTING COLLEGE SCHOLARSHIP.....	357
<i>Stuit, Dewey B.</i>	
THE PREDICTION OF SCHOLASTIC SUCCESS IN A COLLEGE OF MEDICINE	77
<i>Thurstone, Thelma G.</i>	
PRIMARY MENTAL ABILITIES OF CHILDREN.....	105
<i>Traxler, Arthur E.</i>	
CUMULATIVE TEST RECORDS: THEIR NATURE AND USES.....	323
<i>Traxler, Arthur E. and McCall, William G.</i>	
SOME DATA ON THE KUDER PREFERENCE RECORD.....	251
<i>Triggs, Frances Oralind and Sandt, Karl E.</i>	
AN EVALUATION OF TECHNIQUES OF MEASURING VISUAL ACTIVITY AT THE COLLEGE LEVEL.....	295
<i>Tyler, Ralph W.</i>	
CONTRIBUTIONS OF TESTS TO RESEARCH IN THE FIELD OF STUDENT PERSONNEL WORK.....	133
<i>Williamson, E. G. and Bordin, E. S.</i>	
THE EVALUATION OF VOCATIONAL AND EDUCATIONAL COUNSELING: A CRITIQUE OF THE METHODOLOGY OF EXPERIMENTS....	5
<i>Williamson, E. G. and Bordin, E. S.</i>	
AN ANALYTICAL DESCRIPTION OF STUDENT COUNSELING.....	341
NEW TESTS	199
MEASUREMENT ABSTRACTS	96, 205, 311, 409
MEASUREMENT NEWS	101, 318

PRESENTING A NEW JOURNAL



The interest and activity in the field of the measurement of human characteristics have never been greater than today. It is with this thought that the editors present the first issue of *Educational and Psychological Measurement*. Educational institutions, government, and industry are all giving increasing attention to methods of evaluation aimed at determining the status and promise of the individual. Improved methods in measurement are being developed and significant research is being done in many fields.

In spite of this rising interest, measurement is still a stepchild. The contributions of measurement theory and practice have found expression in the publications devoted primarily to other fields. Nowhere has there been a common meeting ground for the exchange of ideas from area to area except for the more technically inclined.

Yet there are measurement problems of practical and immediate concern which are common to many fields. The problem of estimating future success is common, for example, to the tasks of helping young people choose appropriate vocations, to the tasks of helping employees, of admitting students to educational institutions and assigning draftees to jobs in the Army.

The limited interchange of ideas and techniques in measurement probably can be explained by the fact that there has been no single journal which could be counted upon to report current developments and to serve as a forum for the discussion of prob-

lems. It is our purpose to remedy this situation. The pages of *Educational and Psychological Measurement* will be open to contributions from all fields in which techniques of human measurement are used. Each issue of the journal will have departments devoted to news and abstracts of recent literature. Future issues will also carry a section on new tests.

It is hoped that the articles in the journal will not only be of interest to readers in the specific areas from which the articles come, but that they will be suggestive of improved procedures elsewhere.

Washington, D. C.
December 23, 1940.

G. F. K.

THE EVALUATION OF VOCATIONAL AND EDUCATIONAL COUNSELING: A CRITIQUE OF THE METHODOLOGY OF EXPERIMENTS*

E. G. WILLIAMSON AND E. S. BORDIN
University of Minnesota

With increasing attempts to systematize the concepts of counseling, to describe its techniques, and to delineate its objectives, the need for evaluative studies has become more insistent. Descriptions of programs of vocational and educational counseling usually close with a summary statement that further improvement in this field is dependent upon evaluative studies (40: chap. XXVII, 42, 43: chap. IX, 44†). In other words, currently used techniques of counseling must be subjected to scrutiny and evaluation in order that more effective ones may be developed. Thus a fertile field for experimentation may be found in this phase of student personnel work.

Restricting Conditions

A review of the peculiar conditions of this field of applied psychology is in order and should precede attempts to experiment. This paper will attempt to summarize, in a critical and systematic manner, the assumptions, criteria, methods of measuring outcomes, and possible experimental designs involved in the evaluation of educational and vocational counseling. The treatment of personality, social, family and other types of students' problems will be considered only in relationship to educational and vocational adjustment. The evaluation of these other types of counseling—usually called personality counseling—should be the subject of another paper.

When we speak of counseling, we refer to individualized efforts to help students discover vocational assets and disabilities and to plan an appropriate training program. The making of such an inventory of potentialities must be preceded by the collection and use of evidence of abilities, interests and motivations. The techniques involved in collecting, refining and using evidence have been described elsewhere (40: chap. III).

*The report of a statistical evaluation of clinical counseling by the same authors will appear in the next issue of this journal.

†See pages 22-24 for references.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

For purposes of evaluation experiments, it is necessary also to agree as to what counseling is not, to define it negatively. For example, we cannot accept the assumption that testing alone or that statistical prediction is counseling. Such would seem to be the assumption of Thorndike's 1934 study (35) as an evaluation of counseling techniques. On the other hand, attempts to define counseling as self-analysis (by students) or as diagnosis based alone upon impressions, student hopes and interview data (by counselors) are equally unacceptable. Counseling must be based upon an understanding of the student; but the counselor does more than make a diagnosis or prediction. Counseling is the process of helping the student to plan, and to utilize his assets.

Progress toward adequate evaluation of counseling has been impeded by two types of attitudes held by some personnel workers. Some counselors evaluate by means of arm-chair methods. That is, the effectiveness and general worth of counseling is held to be self-evident. These persons reason that the general methodology of guidance must be effective because it appears to be an appropriate method of dealing with serious and widespread maladjustment among youth. Other personnel workers appear to believe that counseling cannot be evaluated. They maintain that the counseling process is so personal and individual that any attempt by the counselor to study it will impair his efficiency as a counselor and will create an artificial situation which will not even remotely resemble the real counseling relationship.

On the other hand, those who believe that counseling can and should be evaluated have taken one of three approaches. First, there is the approach which clings to traditional statistical methodology in utilizing only those criteria that are objectively quantifiable. This approach is based upon the premise that a straightforward statistical analysis of such data as grades, years in college, number of jobs held or wages earned, are sufficient criteria for evaluation experiments. Second is the approach which utilizes non-statistical case study methods of evaluation. The third approach attempts to avoid the objections to the other two methods by using various objective and systematically derived criteria which are combined by means of impartial judgmental treatment in contrast with statistical summations.

The assumptions underlying criteria should be made explicit. Implicit assumptions have been the source of error in planning and interpreting some evaluation studies. For example, prediction has frequently been treated as though it were the beginning and end of

VOCATIONAL AND EDUCATIONAL COUNSELING

guidance. Here again the interpretations of Thorndike's study (35) serve as an example, although others might also be used. The conclusion, drawn by many from Thorndike's study, that counseling was low in effectiveness, would not be objectionable if the interpreters had indicated that by guidance they mean statistical prediction of fragmented criteria. In speaking of prediction of fragmented criteria we refer to the fact that many research workers lose sight of the possibility that one datum often has different meaning and significance for different students. If such is the case, and we have every reason to believe that it is, then any attempt to use these bits of information either separately or in a rigid arithmetic combination may obscure the actual outcomes of counseling.

The supposition that specific objectives, such as an increase in academic achievement, will necessarily be common to all the cases in an experimental population must also be examined. If we cannot accept the supposition, then we must consider the possibility that the use of what is at best a partially applicable criterion is likely to reveal only slight differences, if any at all. For example, a low aptitude student who had been successfully counseled into withdrawing from college cannot be included in an experiment designed to reveal the effectiveness of counseling in increasing grades.

There are two other considerations of this type that the careful research worker must consider in planning an effective evaluative experiment. First he must realize that in order to evaluate a program of action, it must be carried out. The student must do something following counseling in order to make evaluation possible. A physician might just as well attempt to discover the effectiveness of his medicine when his patient has taken it home and placed it unused in his medicine cabinet. Secondly, a counselor may change a student's attitudes, but these must be revealed in observable or measurable behavior or they cannot be evaluated. Any outcome that is beyond the scope of some means of dependable observation is one that cannot be dealt with and therefore must be rejected by those who require more than blind faith.

The question of the optimum time interval for evaluation is one that needs further investigation before much progress can be made in evaluation experiments. It is possible that the optimum time interval will vary for each individual in any experimental group; or perhaps the longer the intervening time, the greater the possibility for the intrusion of other influences that may tend to mini-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

mize the effects of counseling. Some influences may facilitate adjustment subsequent to counseling; others may cause maladjustment. Even though counseling results in a distinct separation of counseled from non-counseled in terms of subsequent adjustment, the randomization of subsequent influences may cause a regression toward the mean for both groups.

The scant knowledge of specific counseling techniques has forced us to study the effectiveness of the total process. If certain techniques neutralize the effect of others, then the gross results would be negligible. As specific techniques are isolated and described, then new types of evaluation studies may replace the present gross experiments. Such studies, however, would not appear to be possible until more adequate descriptions of techniques are made available by those who actually counsel students.

Formulating Hypotheses

Counseling can be evaluated only if certain outcomes or criteria of effectiveness are assumed to result from the counseling process. These assumptions must be formulated as hypotheses to be "tested" by experimental and statistical analyses. But a second consideration is of equal importance. We must determine not only the results of counseling but, as in all scientific studies, the conditions under which these outcomes will be produced. We must answer this second question in terms of what kinds of counseling, what techniques, what types of counselors and work with what types of students will produce certain outcomes. Our problem, broadly speaking, then becomes, "*What counseling techniques (and conditions) will produce what types of results with what types of students?*"

Most counselors have empirically derived opinions, hunches and judgments as to what outcomes or effects they and the students are trying to achieve. But many of these outcomes are intangible and difficult to formulate as well as difficult to set up in an experimental design. We may, however, achieve some degree of agreement, for purposes of experimentation, on the following assumptions:

Effective counseling will lead to or result in:

1. Occupational orientation—understanding and acceptance (choice) of a tentative and broad goal and of the educational (training) means to that goal.
2. This goal will be appropriate to the student in that it will be one which will utilize his aptitudes and interests and will not demand either less or more (within a reasonable range) aptitude than he possesses (actually and potentially).

VOCATIONAL AND EDUCATIONAL COUNSELING

3. The student will make reasonable progress toward this goal (in training school).

4. The student will be "satisfied" (further motivated) by that progress and with his chosen goal.

In order to achieve these outcomes it is necessary that:

1. The counselor secures the student's cooperation (rapport in the broad sense) in choosing (orienting himself toward) a goal and the means to it; the desire to assay his assets and interests.

2. The student generates enthusiasm to use his assets in attempting to secure relevant training and to achieve the chosen goal.

3. The student uses his aptitudes skillfully in securing training in school.

4. The counselor and the student are able to alleviate, relieve or remedy pressures and disabilities—family, financial, emotional, etc.—which interfere with or prevent the eager and skillful use of aptitudes and the choice of an appropriate goal.

5. If these pressures or disabilities are too serious for the counselor to cope with, then use is made of specialized personnel workers.

6. The appropriate or reasonably approximate type of training is available to the student.

The above possible outcomes may be the direct or indirect, immediate or long-term outcomes of counseling. They may reveal themselves or be observed indirectly and not always by means of the student's verbal report to the counselor. For example, the student's orientation may be revealed in his classroom grades. Some outcomes may be general in nature (results of any type of counseling technique), and others may be highly specific. Likewise some techniques may produce one or more of the above outcomes when used with any type of student having any type of problem. Other techniques may be highly specific. Much experimentation needs to be done before we can answer these subsidiary questions. It is most likely that counseling cannot be equally effective with all types of students and all types of conditions.

Experimental Designs

Drawing upon empirical knowledge, we may describe the general outlines of a number of possible experiments which should reveal some of the outcomes of counseling. We shall restrict ourselves to the following possible criteria: academic achievement, appropriate choices, cooperation, satisfaction, success, quality of case work, predictive efficiency, composite criteria.

Academic Achievement. The emphasis placed upon grades in

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

educational circles has necessarily established them as the most used criterion of the effectiveness of counseling. Most colleges and universities drop unsatisfactory students on the basis of their average grades, and reward those who achieve high marks. There are two methods of experimental design applicable: the comparison of the student's grade average before and after counseling (2, 4, 8, 19, 24, 43: chap. IX); or a comparison of the average grade of counseled students with that of non-counseled students who have been matched for such characteristics as age, sex, level of ability, size and type of high school and high-school grades (13, 15, 20, 27, 41, 42).

Both methods of control have definite weaknesses. First of all it must be emphasized that grades are patently only one of the possible desirable outcomes of counseling. In addition their reliability and validity, as a measure of scholastic achievement, have been seriously questioned. Of more importance are the dissimilarities in patterns of subjects taken by different students. This condition makes the criterion of average grade a shifting scale whose comparability from student to student is questionable. Moreover, in cases where the student has been successfully advised to leave college there will be no subsequent grades to evaluate. In the case of students counseled before matriculating in college, where no pre-counseling grades are available, this method is not at all applicable.

The method of control by matching is a traditional one in scientific experimentation. It theoretically provides us with a comparable population for comparing the effect of counseling with the effect of "normal" (or random) conditions. At the present time, however, it is impossible to match individuals on the very factors that may be of importance, e.g., motivation, personality or emotional stability. In addition, it is difficult to collect a reasonable number of cases which will be matchable. While the method of internal control, i.e., comparing grades before and after counseling, does away with the matching problem, it leaves indeterminate the problem of the effect of "normal" conditions in comparison with the effect of counseling.

The use of standardized achievement tests is a possible alternative to grades as a measure of academic achievement. Such tests would be more reliable and presumably would provide a more comparable measure from student to student or group to group. This would be true in any one area of information but, where achievements in a number of areas are to be combined, heterogeneity will again be introduced. As long as individuals differ in the patterns of

VOCATIONAL AND EDUCATIONAL COUNSELING

their objectives and college subjects this factor of heterogeneity will be a possible disturbance in the use of scholastic criteria of counseling effectiveness. Experiments should be made to determine the possible relevancy and validity of this type of criterion.

Educational and Vocational Choices. When evaluating in terms of educational and vocational choices it may be assumed that the individual will achieve a more satisfactory life adjustment if he sets goals for himself that are neither too high nor too low for his potentialities (18, 32). Thus the task of the counselor is conceived to be, in part at least, to bring about congruence between those two factors.

For any case we may compare the student's statement of his objectives with his potentialities as judged from test data and relevant tryout experiences. The judgment of the degree to which better alignment has been achieved as a result of counseling may be made by the counselor himself, by an outsider who reads the case notes, by the student, or by all three persons. In favor of the former procedure, one may contend that there are often subliminal data not included in the case record which would make the counselor's judgment most accurate. On the other hand, we may encounter difficulty in separating judgment from desire since the counselor is not disinterested in the outcome. An added difficulty with this type of criterion is the frequency of student cases in which a temporarily uncertain choice is the most desirable outcome of counseling.

An indirect measure of this criterion may be used if we assume that more information on educational and vocational topics will lead to a greater probability of congruence between aspirations and potentialities. It seems legitimate to expect the clinical counselor to aid the student in acquiring such information, although this type of function has usually been involved in group guidance procedures. For the appraisal of these two types of outcomes, tests and inventories of the Kefauver-Hand type may be used (17). By these means it may be possible to determine whether counseled students have more information on which to base their educational and vocational decisions than they had before counseling or than is possessed by a matched uncounseled control group. Since the mere possession of occupational and educational information is not a major objective of counseling, experiments are needed to determine the relationship between the possession of such information and the appropriateness of the choices made by students. Such crucial experiments have not yet been made in support of the relevancy for counseling of courses in occupational information.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Cooperation with the Counselor. This criterion is based upon the premise that effective results of counseling cannot be achieved unless the counselor is *en rapport* with the student. The fact that the student receives the advice of the counselor cooperatively is taken as an indication of rapport and therefore as a criterion of the effectiveness of counseling. Viteles seems to go even further when he says: "That advice is followed is probably in itself an evidence of satisfactory adjustment" (38: p. 75). Such a contention needs to be evaluated experimentally. We would reject any attempt on a physician's part to prove the efficacy of a particular medical treatment by means of evidence that the patient cooperates in submitting to that treatment. We would certainly withhold judgment until we ascertained whether his patient eventually had recovered or died. Cooperation is a desired outcome of counseling but *chiefly* as a means or condition necessary to other more basic outcomes. In this sense it is a preparatory outcome or criterion of counseling effectiveness.

The measurement of this criterion would be expressed in terms of the percentage of the group counseled that had shown various degrees of cooperation. Such a result is difficult to interpret since there is no standard for determining what either a statistically or a socially significant percentage would be. Further experimentation and experience would, of course, provide data for deriving such a standard.

The Student's Satisfaction. Satisfaction of the student is deemed to be a desirable outcome of counseling. This satisfaction may embrace his educational and vocational objectives, the counseling assistance, and finally the job that he ultimately secures. The student's satisfaction with any of the three may be inferred from his verbal report, either on an interview basis or by means of an attitude test. Obviously many subtle or delayed satisfactions may not be readily observed or felt by the student. Dissatisfaction which results from frustration may be, and oftentimes is, followed by later reconciliation to substitute adjustments.

Concerning satisfaction with educational and vocational objectives as criteria, two methods of control may be used. The satisfaction of the student may be measured before and after counseling or the satisfaction of a counseled group may be compared to that of a non-counseled group. In the case of satisfaction with counseling assistance (25, 39) neither of these methods is possible. To measure a student's satisfaction with counseling assistance before he has been counseled or when he has not been counseled would be mean-

VOCATIONAL AND EDUCATIONAL COUNSELING

ingless. We can only determine the percentage of students who expressed degrees of satisfaction with the counseling assistance received and compare the results for two or more counseled groups. In a sense this criterion is usable to determine which of two or more counseling methods, or counselors, is more effective.

The systematic and quantitative data provided by the attitude scale technique have not as yet been exploited in the evaluation of counseling. There are three types of attitude scales that may be used. First, a scale measuring the student's attitude toward the school and his educational training. Bell has already described such a scale for high school students (3: p. 117-23). Second, a scale measuring the student's attitude toward his vocational objectives. Remmers has constructed such a scale and has used it in the appraisal of the effectiveness of group guidance (28, 29). Third, a scale measuring the student's attitude toward the counselor and the counseling assistance. This type of scale has had practically no application. In fact we have found only two instances of its use reported in the literature (14, 23). The usual approach has been through the report of the individual to direct questioning.

While the student's report is the easiest way to determine satisfaction and cannot be ignored as one type of satisfaction response, it has many weaknesses. For example, it may conceal real dissatisfaction behind a rationalization process. It may be a reflection of dissatisfaction in some other area than education or vocation, e.g., social, recreational, sex. The desire to please the counselor because of fixation or gratefulness may lead to a report of satisfaction. In some cases it seems too much to expect a feeling of complete satisfaction even with the most successful counseling, since a counselor cannot be expected to overcome the false hopes of a lifetime in a relatively short period of time. If the individual's stratum of society requires a level of aspiration far beyond his capabilities, the counselor cannot be expected to bring about complete and immediate satisfaction.

Satisfaction with a job has been the most frequently used criterion of the effectiveness of vocational counseling (4, 5, 6, 16, 22, 23, 25, 30). In addition to the direct report of the student, scores on the Hoppock Job Satisfaction Blank and the number of voluntary shifts in jobs have been used as measures of job satisfaction. All of these criteria lend themselves to the use of both an internal and a matched control. But many objections are encountered to satisfac-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

tion with the job, measured in any manner. Job dissatisfaction may reflect dissatisfaction with the low starting salaries which are characteristic of most jobs rather than with the occupational choice resulting from counseling. The dissatisfaction may also be caused by local conditions on the job, e.g., an unpleasant supervisor, uncompanionable workmates, instead of maladjustment to the work involved. Likewise there are special objections to the use of shifts in employment as a criterion, since it is difficult to distinguish a voluntary from a forced shift. A shift, as measured, is an all-or-none process and time does not allow for a measure of degrees of satisfaction or promotion.

The method of internal control with job satisfaction as the criterion is different from the one previously outlined. In this case before-after comparisons are not applicable. Instead, those who are in an advised occupation are compared to those who are not or with those who are in an occupation not discussed with the counselor. For this control to be meaningful the categories of occupations must be broadly interpreted according to their general functions.

Success on the Job. This criterion assumes that effective counseling should lead students to seek and secure jobs in which they can be successful. It can be measured by employer's reports, number of advancements, number of forced shifts and wages earned. The controls applicable are the same as those for the criterion of job satisfaction (4, 9, 16, 22, 25, 30).

The use of a success criterion has at least four general weaknesses. First of all, success is a relative matter, relative to the student's ambitions and to the reactions of his social group to his achievements. Secondly, success may come years later with many other factors, unrelated to the original counseling, intervening to cause it. Success in school is a more immediate adjustment the student must make before the vocational adjustment is necessary. Thirdly, some students advance vocationally more quickly because of aids from parents or friends and not because of counseling. Finally, this criterion is complicated by the influence of the quality of placement work in the senior year of training and is only remotely a criterion of counseling in the freshman year.

Each of the methods of estimating this criterion has been seriously criticized (33, 34). Employer's reports may be subject to error because of the influence of an adverse personal relationship

VOCATIONAL AND EDUCATIONAL COUNSELING

between employer and employee unrelated to quality of work, because of the state of the labor market or because of atypical successes or failures at the time of the follow-up interview or questionnaire. Quite often there will be problems of locating the employer, especially when the student has experienced a number of shifts within a short period of time (9). The absence of standards for comparison and the difficulty in securing cooperation are also contributory factors to the unreliability of employers' reports.

Number of advancements in employment may be unsatisfactory as a criterion of success because the best occupation for an individual may be one in which there are few opportunities for advancement. In addition, in most cases advancement occurs over a long period of time. The longer the intervening time, the more difficult it is to determine whether the original counseling has been the decisive factor rather than any of the many intervening influences. Likewise, the number of shifts in employment presents drawbacks because of the difficulty in distinguishing voluntary from forced shifts and the all-or-none nature of shifts in jobs.

Paterson and Darley (26: p. 19) and, more recently, Lurie (21) have presented evidence which indicates that shifts in jobs may not always be reliable indices of the individual's adjustment. The older study found that the number of job changes did not discriminate workers unemployed early in the depression from those unemployed late. Lurie found that workers discharged during retrenchment were, as a group, as capable as those retained.

In order for comparisons on the basis of wages earned to be meaningful, it is necessary to compare individuals who are working on jobs where comparable wage scales prevail, a difficult task. Another objection is that wages may reflect extra-individual conditions beyond the scope of the counselor's function.

Quality of Case Work. The type and appropriateness of the various procedures and techniques used by the counselor are assumed to be the marks of good counseling. Studies using such criteria are, however, to be considered as preparatory to final studies of the effectiveness of guidance. It should be recognized, however, that unless thorough-going methods are used there is little point in making an experimental evaluation (42).

A critical analysis of the techniques used by the counselor and a critical reading of case history and interview notes are the most feasible methods to determine their appropriateness (7, 39).

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

An unbiased but well-informed "outside" judge would seem to be the most desirable agent to perform such an analysis of case records. There is a difficulty here in that a well-informed judge is likely to be one who has had counseling experience himself and is, therefore, unlikely to be free of convictions. This method is at best a rough measure of whether the counselor used the particular techniques judged appropriate by other counselors. No measure of the effectiveness of these techniques results from the use of this criterion.

Predictive Efficiency. The efficiency of educational diagnosis by the counselor may perhaps be studied more accurately. One possible experimental setup would compare the efficiency of prediction for pre-college cases by the counselor to that of a statistical predictive equation (5). The problem could be further differentiated by comparing predictions made by the counselor on the basis of preliminary information, tests, questionnaire information and preliminary interview, with predictions after the first counseling interview. This would serve to determine the relative importance or validity of the information and impressions collected in the counseling interview with case data, such as test scores, available to the counselor before he confers with the student. Another differentiating study would involve having a case reader, who had no counseling relationship with the student, predict educational achievement on the basis of all the information available up to, but not including, the counseling interview itself. Such predictions may be compared with those made by the counselor after he interviews the student. Such crucial experiments are needed; a preliminary one will soon be reported by the authors.

There are two assumptions that may be applied here as a basis for evaluating the counseling program. One objective that may be assumed for a counseling program is that of enabling students to compensate successfully for their disabilities in order to succeed. If that is an objective, then the expected evidence of efficiency in the counseling program would be a lower prognostic efficiency of a test battery for counseled students than for non-counseled students. If counseling is effective in this sense, then students who, if left alone, would fail, may succeed.

Another objective of counseling may be to bring all factors other than those of aptitude (interest, opportunity, working conditions and so on) to a common level. Thus the performance of the students would be distributed according to their levels of

VOCATIONAL AND EDUCATIONAL COUNSELING

ability. The greater the excess of predictive accuracy for counseled over non-counseled, the closer the counseling program will be presumed to have come to the ideal—that of removing all influences other than ability which interfere with student achievement.

Composite Criteria. All of the criteria discussed above have been partial criteria, since none of them was assumed to be evaluating all the possible objectives of counseling. We turn now to possible methods by means of which a more comprehensive evaluation of counseling may be secured.

The Use of a Judgment Criterion

It is at this point that a clear schism appears between an approach which is narrowly statistical and an approach which makes use of statistical methods in conjunction with the experimental situation. The former point of view has the desirable objective of clear-cut results, but, in its blind adherence to traditional method, produces results which are unlikely to be significant either statistically or socially. This method would mechanically pool all of the part-criteria either in some form of average or in a profile. The method of averages compounds the artificiality which previously had been indicated as inherent in the use of the part-criteria without reference to the individuality of each student. The method of profiles suffers from a lack of well developed statistical techniques for handling that type of data and, more seriously, from the fact that artificial data cannot be refined and validated by casting them into profile form.

Rather than sacrifice meaningfulness for neatness of statistical treatment, the other approach has clearly recognized the impracticability, at the present, of getting more than rough measures of the general efficiency of counseling (33, 37). It has therefore attempted to use a *judgment criterion* by means of which the adjustment of the student is *estimated* in terms of his original problems and any of the available data, including the part criteria (16, 22, 27, 30, 31, 36, 38, 42, 43: chap. IX).

As described in Williamson and Darley the judgment of adjustment is based upon a follow-up interview of the student (43: chap. IX). The status of the case at the time of follow-up is always considered in the light of the diagnosis and prognosis made earlier by the counselor. All of the various types of data—grade achievement, the student's statement of satisfaction and adjustment with regard to vocational orientation and choice,

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

information concerning the student's activities, judgment of general attitude, etc.—are weighed by the judge according to their relevance to the individual case. In this way the possible errors inherent in a non-personalistic interpretation of objective data are minimized.

The simplest experimental design requires that either the counselor or a case reader make the judgment as to the degree of the student's adjustment subsequent to counseling and in contrast with his pre-counseling adjustment. A detailed manual of directions, including examples of degrees of adjustment, is necessary in such an experiment (43: chap. IX). The results are reported in terms of the percentage of students counseled who achieved various degrees of adjustment. If the counselor makes such a judgment, it should be pointed out that, if he is a good one, he will know subtle angles and attitudes which are unlikely to be explicitly stated in the case records and which would be overlooked by an independent case reader. Many of the subtle influences in a case may even exist in un verbalized form for the counselor and have no possibility of appearing in the case record. In addition, the counselor knows, perhaps better than anyone else possibly can, what he has been trying to do. The disadvantages of using the counselor's judgment lie first of all in his special interest in the results which may lead to an approach which is either too self-critical or too self-lenient; and secondly, in the undesirable consequence that the counselor's effectiveness in counseling may be decreased because of his awareness of his responsibility for evaluating his own efforts.

While the use of the independent case reader obviates the possibility of impairing the counselor's effectiveness and introduces a theoretically impartial evaluator, it also has its drawbacks. As has been indicated, the case reader may miss many of the nuances. There are also so many conflicting philosophies and procedures and techniques in counseling that the case reader may be either unsympathetic with or ignorant of the counselor's specific objectives. In order to achieve greater impartiality and objectivity, two case readers and an arbitrator have been used. Williamson has reported the use of this method with three trained workers who had nothing to do with the diagnosis and counseling, but who collected data directly from the students for independent and pooled judgments of effectiveness (42). With trained judges heterogeneity of point of view need not interfere with consistency in judgments.

VOCATIONAL AND EDUCATIONAL COUNSELING

While the experimental method outlined may yield evidence of the degree of improvement in the counseled population, it does not prove that cooperation with the counselor's suggestions was a necessary condition. Evidence for the latter may be obtained by estimating the degree of cooperation of each student by one of the methods previously discussed. By comparing the adjustment achieved by students who cooperated, with the adjustment of those who did not, we can determine whether cooperation was necessary and to what degree. If we find that those who cooperate adjust better than those who do not, we still have a question of whether the degree of adjustment achieved by those who did not cooperate might not be equalled or bettered by those who received no counseling at all. A matched non-counseled group would seem to be the only means of providing an answer to this question. We have already discussed the possibilities of the matching process. If we were to attempt to avoid the matching problem by counseling every other student who comes for counseling, retaining the other half of the group as controls, we would be doing violence to a social canon. The real solution must await a time when we have sufficiently isolated treatment techniques and problems to compare two treatments used with the same type of counseling problem.

General Considerations

Our consideration of the types of criteria and the methods of measuring them, feasible in the evaluation of the effectiveness of counseling, has touched upon definite limitations on exact evaluation of counseling. Whether these weaknesses will be insurmountable and will restrict evaluation to rough, rule-of-thumb methods depends upon future progress in experimentation.

One type of difficulty is the inability to set up clear delineations of the problems and variables involved. This has been traced first to the inadequacy of descriptions of diagnostic and treatment techniques of the counselor plus the gaps in knowledge of student problems (40: chap. XXVII). A second source is the element of uniqueness in the student's problems and the counseling techniques appropriate for them. The criteria which have been considered have the weakness of being either too gross a measurement or so far removed from the individual as to lack the quality of meaningfulness. If, in the future, methods are devised for providing more adequate criteria, then more exact experimentation may be made.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

A second limitation, inherent in the nature of the counseling situation, is the sampling difficulties involved in setting up controls. This difficulty has been stated by Murphy *et al.*. "In connection with the control group, there should be noted the relative impossibility of securing a true 'control.' The fact that the experimental group applied voluntarily for counseling introduces a selective factor unmatched among non-clients" (23: p. 952). Although this assumption has never been established *even in the plethora of learning experiments*, only assumed, caution should be applied in interpreting the results of matching experiments. We also lack adequate techniques for matching students for such pertinent factors as interest, perseverance and other similar qualities. At the same time, it is impossible to set up an experiment which would entail selecting cases from the general population, since willingness to be counseled would seem to be one of the necessary conditions for counseling. These limitations in sampling methods imply that evaluation must necessarily be a long-time process, involving a great deal of experimentation with different methods.

The condition that diagnosis and counseling cannot be studied separately is a further complicating factor. When the counselor has made a diagnosis of the student's problems, its causes, and the types of treatments that are likely to solve it, he cannot determine whether his diagnosis was correct unless the student carries out the recommendations. For example, if a counselor's diagnosis states that student A can do effective work in college only by following certain of his recommendations, student A must remain in college for that diagnosis to be tested. The inability to control the conditions necessary for an adequate tryout of counseling recommendations often precludes determination of the effectiveness of the advice. Factors which are often beyond the control of either the counselor or the student include restriction imposed by the school administration, those imposed by social codes, prejudices and attitudes of students or parents and lack of proper placement facilities.

The criteria and methods discussed in this paper have little application for the comparison of individualized counseling with other types, e.g., group, traditional, casual interview, etc. This situation arises from the nature of the data yielded from these kinds of counseling. For example, the casual interview, by its very nature, does not produce very much information about the individual's aspirations, his difficulties, or the counseling methods

VOCATIONAL AND EDUCATIONAL COUNSELING

used by the counselor. Grade or information achievement represents the only type of criterion that can be applied in evaluating this type of counseling.

Our discussion has shown that there is a need for more systematic studies, using the more feasible part-criteria. Other approaches have been indicated as having possibilities. The relationship between the student's educational and vocational objectives and his level of ability should be studied further as a criterion of adjustment. Some studies have already yielded some preliminary results (1, 10, 29, 32). In the last few years this problem has been receiving attention under the term level of aspiration. Studies have revealed some provocative principles under laboratory conditions (11, 12), but we must learn whether these principles have validity for life situations. We should determine whether success in one area, i.e., vocational or educational, has an effect on the level of aspiration in other areas of adjustment. What are the relations between level of aspiration and feelings of failure? Can vocational or educational success be such a potent factor that it would outweigh other experiences in determining an individual's general success-failure feelings? How vital are social group factors in determining the individual's level of aspiration? To what degree do levels of aspiration persevere at various age levels? The answers to these questions would seem to be pregnant with implications for both the counselor and the evaluator.

If and when our knowledge of student problems and of diagnostic and treatment techniques has advanced sufficiently, we will have the opportunity to carry out more exact investigations. At this point, we can foresee experimental designs which should be applicable when such advances are made. One possibility is an experiment in which individuals having problem A will be divided into two groups, one which will receive treatment 1, the other treatment 2. In this way we may determine which specific techniques are most effective for a particular problem.

Another plan of experiment could be designed to determine for what types of problems a treatment is applicable. Here, two groups, one representing problem A, the other problem B, would both receive treatment 1. Both methods could be expanded to include all types of treatments and problems. Criticism of these designs may be directed at the apparent assumption that problems may appear isolatedly. That this is extremely unlikely cannot be denied. Yet, assuming advances in our techniques and proced-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

ures, it seems possible that the types of factorial design used in analysis of variances can be utilized to take care of the effects of interactions among treatments for various problems. The success of such an experiment will depend upon the discovery of a number of cases in which one problem is clearly present and other types are minimal in significance or complexity.

Summary and Conclusions

1. All available methods of evaluation have weaknesses.
2. Composite criteria which avoid arithmetic combination of the part-criteria are at present least open to question, although still being crude measures.
3. The problem of securing sufficient data without doing violence to the concept and practice of counseling is a real one. Involved also are the inadequacy and incompleteness of most available case records.
4. The proper time interval to use for evaluation is extremely important because of the possible relationship between the intervention of confusing factors and the length of time between counseling and evaluation.
5. The methods used for validation of diagnostic and prognostic tools (e.g., tests) may not be applicable because of the uniqueness of each counseling situation. Stated another way, the methods of studying students in general may not be applied to the study of individual students with particular problems.
6. An impediment to more exact evaluation is the inability to control conditions for an adequate test of counseling recommendations.

BIBLIOGRAPHY

1. Alberty, H. B. "The Permanence of Vocational Choices of High School Pupils." *Industrial Arts Magazine*, XXIV (1925), 203-07.
2. Beaumont, H. "The Evaluation of Academic Counseling." *Journal of Higher Education*, X (1939), 79-82.
3. Bell, Hugh M. *The Theory and Practice of Personal Counseling*. Palo Alto: Stanford University Press, 1939.
4. Burt, Cyril & Others. *A Study in Vocational Guidance*. Industrial Fatigue Research Board. Report No. 33. London: H.R.H. Stationery Office, 1926.
5. Clark, E. B. "Value of Student Interviews." *Journal of Personnel Research*, V (1926), 204-07.
6. Cole, R. C. "Evaluating a Boys' Club Guidance Program." *Occupations*, XXVII (1939), 705-08.

VOCATIONAL AND EDUCATIONAL COUNSELING

7. Coler, C. S., Fitch, John A., Fitch, Florence Lee, Paterson, Donald G. *General Appraisals of the Adjustment Service*. New York: American Association for Adult Education, 1935. 87 pages.
8. Cowley, W. H. "An Experiment in Freshman Counseling." *Journal of Higher Education*, IV (1933), 245-48.
9. Earle, F. M. *Methods of Choosing a Career*. London: George C. Harrap & Company, Ltd., 1931.
10. Feingold, G. A. "The Relation Between Intelligence and Vocational Choices of High School Pupils." *Journal of Applied Psychology*, VII (1923), 152.
11. Frank, Jerome D. "Individual Differences in Certain Aspects of Level of Aspiration." *American Journal of Psychology*, XLVII (1935), 119-28.
12. Frank, Jerome D. "Some Psychological Determinants of the Level of Aspiration." *American Journal of Psychology*, XLVII (1935), 285-93.
13. Freeman, H. J. and Jones, L. "Final Report of the Long-Time Effect of Counseling Low-Percentile Freshmen." *School and Society*, XXXVIII (1933), 382-84.
14. Hawkins, L. S. and Fialkin, Harry N. *Clients' Opinions of the Adjustment Service*. New York: American Association for Adult Education, 1935. 95 pages.
15. Holaday, P. W. "The Long-Time Effect of Freshmen Counseling." *School and Society*, XXIX (1929), 234-36.
16. Jennings, J. R. and Stott, M. B. "A Fourth Follow-up of Vocationally Advised Cases." *Human Factor (London)*, X (1936), 165-74.
17. Kefauver, N. and Hand, H. C. *Manual for Kefauver-Hand Guidance Tests and Inventories*. New York: World Book Company, 1937.
18. Kirkpatrick, F. H. "Vocational Guidance in An American College." *Human Factor (London)*, XI (1937), 409-14.
19. Leman, A. C. "An Experimental Study of Guidance and Placement of Freshmen in the Lowest Decile of the Iowa Qualifying Examination, 1925." *University of Iowa Studies in Education*, III (1927), 8. University of Iowa.
20. Lund, S. E. Torsten. "The Personal Interview in High School Guidance." *School Review*, XXXIX (1931), 196-207.
21. Lurie, W. A. "Intra-Individual and Extra-Individual Factors Influencing the Levels of Vocational Aspiration and Achievement." A paper read at the Forty-sixth Annual Meeting of the American Psychological Association, Columbus, Ohio, 1938. Abstract in *Psychological Bulletin*, XXXV (1938), 670.
22. MacRae, A. "A Follow-up of Vocationally Advised Cases." *Journal of the National Institute of Industrial Psychology*, V (1931), 242-47.
23. Murphy, J. F., Hall, O. M., and Bergen, G. L. "Does Guidance Change Attitudes?" *Occupations*, XXIV (1936), 948-52.
24. Newland, T. Ernest and Ackley, W. E. "An Experimental Study of the Effect of Educational Guidance on a Selected Group of High School Sophomores." *Journal of Experimental Education*, V (1936), 23-5.
25. Oakley, C. A. "A First Follow-up of Scottish Vocationally Advised Cases." *Human Factor (London)*, XI (1937), 27-31.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

26. Paterson, D. G. and Darley, J. G. *Men, Women, and Jobs*. Minneapolis: University of Minnesota Press, 1936.
27. Paterson, D. G. and Langlie, T. *Report of a Controlled Experiment on the Value of Faculty Advisers for Probation Students in the College of Engineering, Chemistry and Architecture, University of Minnesota, 1925-26*. (Unpublished.)
28. Remmers, H. H. "Measuring Attitudes Toward Vocations." *Studies in Higher Education, Purdue University*, XXXV (1934), 77-83.
29. Remmers, H. H. and Whisler, L. D. "The Effects of a Guidance Program on Vocational Attitudes." *Studies in Higher Education, Purdue University*, XXXIV (1938), 68-82.
30. Rodgers, T. A. "A Follow-up of Vocationally Advised Cases." *Human Factor (London)*, XI (1937), 16-26.
31. Seipp, Emma. *A Study of One Hundred Clients of the Adjustment Service*. New York: American Association for Adult Education, 1935. 30 pages.
32. Sparling, E. J. *Do College Students Choose Vocations Wisely?* New York: Teachers College Contributions to Education, Columbia University, 1933.
33. Stott, Mary B. "Criteria Used in England." *Occupations*, XXIV (1936), 953-57.
34. Stott, Mary B. "Occupational Success." *Occupational Psychology (London)*, XIII (1939), 126-40.
35. Thorndike, E. L. *Prediction of Vocational Success*. New York: The Commonwealth Fund, 1934.
36. Trabue, M. R. and Dvorak, B. J. *A Study of Needs of Adults for Further Training*. Minneapolis: University of Minnesota Press, 1934.
37. Viteles, M. S. "A Dynamic Criterion." *Occupations*, XIV (1936), 962-67.
38. Viteles, M. S. "Validating the Clinical Method in Vocational Guidance." *Psychological Clinic*, XVIII (1929), 69-77.
39. Williamson, E. G. "Faculty Counseling at Minnesota. An Evaluation Study of Social Case Work Methods." *Occupations*, XIV (1936), 426-33.
40. Williamson, E. G. *How to Counsel Students*. New York: McGraw-Hill, Inc., 1939.
41. Williamson, E. G. "The Role of Faculty Counseling in Scholastic Motivation." *Journal of Applied Psychology*, XIX (1936), 314-24.
42. Williamson, E. G. *A Summary of Studies in the Evaluation of Guidance*. Report of the Fifteenth Annual Meeting of the American College Personnel Association, 1938. Pp. 73-7.
43. Williamson, E. G. and Darley, J. G. *Student Personnel Work*. New York: McGraw-Hill, Inc., 1937.
44. Wrenn, C. G. *Recent Research on Counseling*. Report of the Sixteenth Annual Meeting of the American College Personnel Association. Cleveland, Ohio, 1939. Pp. 88-94.

THE LOGIC OF AGE SCALES*

M. W. RICHARDSON

United States Civil Service Commission

An age scale is a type of psychological test designed to measure general mental ability ("intelligence") in terms of performance of various mental tasks found to be normal for various ages. The child whose performance is typical of ten-year-old children, for example, is said to have a mental age of ten. The method was invented by Alfred Binet. Although the techniques have been modified in detail by British and American psychologists and the device of the intelligence quotient (I.Q.) has been appended, the main outlines of Binet's work have been retained. The most widely used of the age scales is the Stanford Binet.

The age scale has been widely accepted. The I.Q., in particular, has passed into the language of the general public, together with the common misconceptions connected with its brief history in science. The fact that a device has attained wide use is not a guarantee of its soundness; and it is sometimes necessary in the interest of sound scientific advance to examine critically procedures and devices in common use. If the criticisms in this paper seem to be directed chiefly to one particular age scale, the explanation is that this one scale is the most widely used and has been most carefully constructed and standardized.

A person whose academic specialty is the logic of science addressed a group of psychologists on the necessary conditions of measurement. He discussed the familiar matter of equality of units and the operational test of equality of units by the coincidence of any part of the scale with any other upon superimposition. He mentioned the matter of measuring a single variable at a time, and the necessity of having a real origin of measurement, if ratio comparisons are to be made. He followed this sound discussion of the scientific method with a curiously erroneous one; he congratulated psychologists in having, in the Binet age scales, a measuring device that meets all three of the requirements for a scientific measuring device. The writer and others carefully

* This article is adapted from a chapter in a forthcoming book on test theory by the same author.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

pointed out to the speaker that the age scale meets not one of the three requirements set up; it has no real origin of measurement; its units are not equal, and it does not isolate a single, unitary variable for measurement.

When it is uncritically considered, the age scale idea seems to be a happy one. What could be more simple and direct than to see how high the child can ascend an evenly graded scale? The concept of normality of performance made it the most natural thing in the world to describe such test performance in terms of the age for which the performance was normal. The concept of the age scale is a deceptively simple one, however, and the writer is of the opinion that the unstated special requirements and limitations of the age technique have received too little attention. It is true that during the twenty-one years between 1916 and 1937 many papers pointing to difficulties in scaling, scoring, and interpreting the Stanford Binet appeared. Moreover, several issues were kept continually in the foreground of attention. Unfortunately, the distinction between purely psychometric issues and psychological issues was not always made. The result is that certain deficiencies and limitations belonging to the mechanics of test construction were misinterpreted as psychological issues. A case in point is the constancy of the I.Q., about which it will be necessary to say more later.

Validity of Age Scales

A Binet scale consists of a series of sub-tests or items designed to measure "general intelligence," whatever that may mean. For example, the 1937 edition of the Stanford Binet contains 127 sub-tests graded in difficulty from tests suitable for two-year-olds to those suitable for superior adults. The sub-tests were selected in the process of construction from a larger number of sub-tests. It is pertinent to inquire into the method of selection of the sub-tests. In what respect does the method of selecting the sub-tests insure that the resulting scale will be valid? One of the devices is to plot the percentage of correct responses to any given sub-test against the chronological age, after the sub-test has been applied to "unselected" children of various age groups. The age at which just half of the children pass the test is taken as the scale-position of the item. The plots of percentage of correct answers against chronological age differ from item to item; some curves are steeper than others. Theoretically, the items with steep curves are selected to make up the scale; actually, in the construction

THE LOGIC OF AGE SCALES

of the Binet tests many compromises with practical expediency are made. Under special conditions a mathematical function can be used to describe a discrimination curve of items against chronological age. It has been shown that the use of this function is simply an alternative to the correlation methods, and precisely equivalent to them under certain special conditions.

This type of analysis throws light on the "validating" procedure. The retention of sub-tests on the basis of sharp curves of age discrimination is the same as retaining items that have the large correlations with chronological age. The criterion of validity is simply chronological age, and the practical effect of the procedure is to select items that have relatively high correlations with chronological age.

The procedure leads to a serious difficulty. The standing high jump, or other athletic skills, yield similar discrimination functions, since they are likewise positively correlated with chronological age. The method of item selection thus breaks down as a way of attaining validity. The only criterion of validity remaining is the judgment of the persons constructing the scale. An allied consideration is that the selection of items on the basis of high correlation with mental age on the same or previous scale, is merely a measure of internal consistency or reliability. Nothing in the general procedure operates towards the selection of items that measure a unique trait. An interesting logical difficulty appears. Suppose that, out of the hand-picked collection of items supposedly measuring the aspects of intelligence desired in the scale, the items selected are those which have the steepest discrimination functions. Let us assume further that two items are so discriminating and so far apart in proper age-location that their discrimination functions do not overlap. The result is that the two hypothetical "good" items or sub-tests have a zero correlation. A scale made up of such items must necessarily be unreliable as a composite measure. Furthermore, to the extent to which the search for valid sub-tests by this procedure should be successful, the number of different factors measured would increase. Evidence at present suggests that no fewer than six different mental functions are measured in a higgledy-piggledy fashion by the Stanford Binet. The multiplicity of factors is perhaps not so serious as the fact that different things are measured at different ages. The sobering fact about the age-scale technique is that we do not know what is being measured, or what any given intelli-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

gence quotient means in terms of the relative standing of the individual.

As an added complication, the score received by the testee is expressed in terms of *mental age*. Mental ages are statistical numbers based on the concept of normal or average test performance of children of a given age, when the sample of children is representative of some population of children. On page 25 of *Measuring Intelligence*, Terman and Merrill state that the expression of a test result in terms of age norms rests upon no statistical assumptions. The statement is erroneous and misleading. The truth of the matter is that the *mental age* is a measure derived from raw scores in accordance with certain assumptions; it is as definitely statistical in nature as the standard score, for example.

In using scales of the Binet type, we choose to express test performance, not in the arbitrary units of number of items passed, but in terms of "mental years and months." The raw score "units" are of course arbitrary, in the sense that they are not units at all. The child who answers 12 items correctly cannot be said to exceed the child who passes 9 by the same amount that the latter surpasses a third child who answers 6 correctly. No ordinary test can be expected to satisfy the additive property required for measurement on a scale. But the mental year or the mental month is likewise not a real unit of measurement. In order for the mental year (or month) to be a real unit of measurement, it would be necessary for the function representing mental growth to increase regularly with chronological age. If, during each year, a child had the same increment of mental growth, the mental year or mental month would be constant in value. However, it is commonly agreed that the child matures less and less rapidly as he grows older, in intelligence as well as in physical characteristics. The annual increment of "intelligence," i.e., a mental year, steadily becomes less until mental maturity is reached, at which time it is zero. An age scale is, therefore, not a true scale because it is not built up from equal units. In this connection, it may be noted that the true shape of the mental growth curve cannot be determined from scores expressed in terms of mental ages. If such were attempted, one would get results predetermined by the crude growth curve adopted in order to express raw scores in terms of mental ages.

Whatever the merits of the assumed growth curve may be, the crucial consideration is that its true shape and its upper limit cannot be determined by use of a "scale" expressed in terms of

THE LOGIC OF AGE SCALES

mental ages. The true shape of the mental growth function can be determined, strictly speaking, only by use of a scale with real units of measurement. Once the mental growth function is established, it is possible to calibrate the underlying true scale in terms of mental ages, if desired. Then the interval between the mental age of four and the mental age of five might be expressed as a certain fraction of the scalar unit, the mental year between five and six as a somewhat smaller fraction of the same real unit, etc. Finally a place would be reached where the mental year is a negligible fraction of the real unit, and therefore has a value of practically zero. We would then have a proper (although indirect) experimental solution of the problem of the limit of mental maturity. The widespread use of mental ages has not helped to solve the problem, mainly because the use of mental ages as derived measures begs the question.

Although the exclusive use of mental ages forever begs the question of the limits of mental maturity, it is urged that a definite and well-accepted social meaning has been attached to them. It seems simple enough to define mental age as the average or median test performance of typical nine-year-old children. The definition works well enough until the limit of mental maturity is reached. If the limit of maturity is assumed to be 15, a mental age of more than 15 is impossible, by definition. Mental ages of more than 15 are assigned in the process of standardization to test performances by use of "cut-and-try" procedures based on some not well-defined assumptions. At best it is unfortunate that the definition of mental age must be radically shifted at one point or region in the age scale.

The I. Q. and Its Troubles

To multiply confusion, the device known as the intelligence quotient has been adopted. The intelligence quotient is defined as 100 times the ratio of mental age to chronological age, and is thus an index of brightness. An index of brightness can of course be no more than a statistic relating the individual's test performance to the average performance of those of the same age. It is exactly as true of I.Q. as it is of other possible statistics serving the same purpose that one must always use it in connection with some measure of variability of test performance within the age group. Obviously one measure taken from a distribution has no meaning unless a measure of dispersion is given. It might be argued that clinicians keep in mind some kind of subjective scale

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

which serves in lieu of the ordinary statistical parameters. If so, the mental feat is remarkable since the various age groups, in one age scale at least, have dispersions of intelligence quotients which vary considerably. The difficulty of interpretation of the single statistic (the I.Q.) is greatly increased where the standard deviation of I.Q.'s of one chronological age group may be twice as large as that of another age group. The intelligence quotient shares with the mental age from which it is derived the fictitious character of measures above the maturity level. It is nonsense to describe an adult as having an I.Q. of 120 because such a statement is based on an irrational definition and is unverifiable experimentally. It is the practice of Binet testers to assume some definite chronological age as the upper limit of mental growth. Thus, it is assumed on at least two age scales that the upper limit of the average child is reached at the age of fifteen. The crucial difficulty in making such an assumption is not that the upper level set may be wrong, but it lies in the utter impossibility of checking up on its correctness by means of age scales.

One of the moot questions about the I.Q. is its constancy. It seems unfortunate to the writer that so much time of psychologists has been wasted on such a matter. It seems that what ought rightly to be merely a formal problem in test construction has been translated into one of spurious psychological significance. The only question properly asked at this time may be definitely stated: Did the authors of the age scale succeed in constructing a device which gives a constant I.Q.? Questions involving changes in I.Q. possibly attributable to environmental factors must always take into account the fluctuations of the I.Q. which are due to the test and to the statistical operations used to determine the intelligence quotient. Failure to consider the expected magnitude of fluctuation of the I.Q. may easily result in gross misinterpretations of time-changes in its value for any one individual.

Before we can properly evaluate the effect of organic and environmental factors on the intelligence quotient, we are forced to consider the variations in the I.Q. inherent in the testing technique. The fluctuations are those associated with the concept of reliability. For the 1937 Revision of the Stanford-Binet Scale the estimated reliability coefficient varies from 0.90 to 0.98, the higher reliability being associated with the lower I.Q. intervals. A median value for those near 100 I.Q. is 0.92. A representative value of the standard error of measurement is 4.5. Certain systematic variations in the individual I.Q. are also found. The

THE LOGIC OF AGE SCALES

practice effect is one type of systematic variation. Terman and Merrill estimate that the mean increase in I.Q. on the second test (which means on the other form, since two forms, L and M, are provided) ranges from 2. to 4.4, when the time interval between testings is short. The increase due to practice effect is presumably greater when the same form is repeated.

Another systematic source of error may lie in details of the construction of the age scale. Thus, it might be a characteristic of certain age scales that the I.Q. of a superior child decreases with chronological age. Strictly speaking, nothing in the definition of the I.Q. requires constancy during the entire period of development. The constancy of the I.Q., if it exists, is imposed by the process of standardization. Therefore, an experimentally obtained constancy of the I.Q. proves only that the scale has been constructed in such fashion as to produce constant I.Q.'s except, of course, that random fluctuations will still be present. When, and only when, an age scale has the characteristic that I.Q.'s of individuals at all age levels tend to remain constant, one may attach significance to the case of the unusual individual whose I.Q. does not remain constant. The significance of any such shift of I.Q. in time must be judged in relation to the normal shift to be expected by random error, or unreliability.

Increases or decreases of the order of magnitude of three times the standard error of measurement must first be tested with respect to the magnitude of variable errors of measurement before it is legitimate to entertain the hypothesis that some other factor such as special therapy, change in environment, or organic change is responsible for the shift in I.Q. In addition, obtained increases should be scrutinized carefully from the standpoint of possible practice effect. All such interpretation is predicated on the basis of constant I.Q., as built into the scale itself. One may properly inquire just how a scale with constant I.Q. may be constructed. In view of the inherent difficulties with the mental age and intelligence quotient, it is impossible to state any perfectly general rules. However, the part of the scale which is treated as if the growth curve were linear, viz. two to 13 years, can be abstracted for discussion. The problem the test constructor faces is that of providing that most individuals will be assigned the same I.Q., within the reliability of the scale, every year from two to 12 inclusive. If the various sub-tests are properly scaled, i.e., assigned to a given year level as a median performance of unselected children of that age, the I.Q. of 100 will remain constant.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Inconstancy is to be expected in intelligence quotients other than 100, unless certain conditions are satisfied. One condition is that the variability of test performance increases for successive age groups from two to 12. If we assume that children of a given typical group vary more among themselves as they grow older, it is possible to arrange matters so that the I.Q. of most individuals will be constant. Suppose that we have an individual whose I.Q. at age eight is 120. The mental age from which the I.Q. is estimated is 9.6 years (or 9 years, 7 months, approximately). Let us further suppose that we have a distribution of mental ages of all the children in our sample of eight-year-olds. The mental age of 9.6 years is, say, one standard deviation above the mean of the distribution of eight-year-olds. The standard deviation of the distribution is $9.6 - 8 = 1.6$ mental years.

Now, if the same child is to have an I.Q. of 120 at the age of nine, his mental age will then be 10.8 years. If the nine-year-old sample is composed of the same children we should expect, except for errors in measurement, that the child will have the same position in the nine-year distribution that he had in the distribution of eight-year-olds. Since his mental age is now 10.8, the standard deviation of the distribution of nine-year-olds is 1.8 mental year. Similarly, the standard deviation of the ten-year-olds must be 2.0 mental years; of eleven-year-olds, 2.2 mental years; etc. The preceding illustration shows that, for an assumed linear growth function, the standard deviations of the mental ages must have constant increments for each advancing year. How shall this be done? Considering, for sake of simplicity, that we have just six sub-tests at each year level, we may increase the standard deviations of successive year levels by (a) selecting sub-tests which have higher intercorrelations at the older age levels, (b) assigning a larger number of mental months to each sub-test. It will be seen at once that the latter is inadmissible since a total of 12 mental months is assigned at each year level. The conclusion is inescapable that the degree of correlation between sub-tests must increase steadily with higher age levels if the I.Q. is to be constant.

The foregoing treatment is theoretical and does not imply that the authors of any age scale have consciously attempted to attain constancy of the I.Q. in such a manner. More probably they have taken advantage of the fact that variability of mental performance does increase with age. Such increase affects the arbitrary units of measurement employed in somewhat unpre-

THE LOGIC OF AGE SCALES

dictable fashion; it may be sufficient to account for the relative constancy of I.Q.'s attained in age scales.

It is difficult, however, to account for the attainment of constancy of I.Q. and, at the same time, approximately equal dispersion at the various age levels. The range of standard deviations of I.Q.'s reported for various half-year levels is from 12.5 to 20.7. A representative value is 17. The values vary considerably, probably because of accidents of scale construction. Certainly the values given by Terman and Merrill do not vary systematically with age, and the authors assume that the true variability is nearly constant from age to age. However, let us consider a half-year group as an approximation to "point" age. If all individuals within such a half-year group are considered to be of the same chronological age, the mental ages are proportional to the intelligence quotients, i.e., a plot of M.A. against I.Q. is linear. Even for a half-year interval, approximate linearity must hold; otherwise the definition of an I.Q. is meaningless.

It follows that if half-year groups have the same I.Q. dispersion, they must have approximately the same mental age dispersion. But the mental age dispersions must increase from year to year in order for I.Q.'s of individuals to be constant. It thus appears that two possible properties of the I.Q. are inconsistent and not attainable at the same time, in any strict sense. The most serious criticism to be directed against Terman and Merrill's discussion of the matter is that they tend to treat the (roughly) approximate equality of dispersion of I.Q. at the various age levels as experimental facts, as perhaps having psychological significance. The I.Q. is a statistical concept, having the properties we put into it by the accidents of cut-and-try scale construction or which we force it to have by conscious design. If we postulate that our statistical index shall have certain properties, we can then construct a test in accordance with our imposed requirements.

The only reservation is that we may possibly have imposed characteristics which are mutually inconsistent, in which case we perforce discover the source of the difficulty. If we fail to limit the properties of a statistic by rational design, the vagaries of that statistic will be brought to light in subsequent empirical studies. The result is the raising of such false issues as the constancy of the I.Q. The gist of the matter is that the I.Q. can be made to be constant, if that is thought to be a desirable property. If the scale is not constructed in such a way as to give

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

constant I.Q.'s for (most) individuals, the die is cast and the I.Q. will not be constant.

In summary, it may be stated that the age scale technique—

- (1) possesses no advantage over group test methods
- (2) has no straightforward rationale so that the process of standardization may proceed without the necessity for "adjustments"
- (3) meets none of the three requirements of real mental measurement
- (4) leads to much useless work in correcting the previous "standardization"
- (5) embodies the mental age and intelligence quotient, both extremely unfortunate concepts
- (6) leads to problems of spurious psychological significance, such as the constancy of the I.Q.
- (7) makes impossible any solution of the mental growth function
- (8) has led to dubious devices and untenable interpretations of various sorts, among them "scatter" and measure of "mental deterioration."

It is recommended that the age scale technique in its present form be abolished in its entirety, and that it be supplanted by reliable homogeneous group tests of single functions. The latter can be recombined, if desired, into a single index of mental capacity based on position in year group. A better procedure is to continue work towards some real unit of measurement, to the end that departures from normal growth in several functions may be discovered and clinically interpreted. If, by reason of demand from teacher, parent, or psychiatrist it will seem necessary to give a general index of (average) mental level reached, it can be done by use of a suitable combination of measures furnished by the separate tests. It is desirable, however, to avoid the use of a single index of mental level.

It has been urged in defense of the Binet test that during its administration, the trained clinical psychologist has an opportunity to make observations of the child's behavior other than that required for rating "general intelligence." It is maintained that such observations may have as much value as, or more value than, the mental age, in getting a "clear picture of the individual tested." If such clinical insights can be reliably obtained and recorded, the obvious desideratum is a standardized interviewing technique, to be applied and interpreted entirely separately from the measures of primary abilities.

COUNSELING ON THE BASIS OF INTEREST MEASUREMENT*

JOHN G. DARLEY
University of Minnesota

As the counselor studies his available data on abilities, achievement, interests, personality, and background of the student facing him in the interview, he must select a conversational starting point that will establish rapport and get the interview under way. At some early time he must discuss the student's stated reason for seeking help, and eventually he must interpret the interest test data in a manner understandable to the student. Assume that the student makes A scores on the occupational keys for Y.M.C.A. secretary, and personnel manager, and B: for school superintendent and social science teacher on the Strong Vocational Interest Blank. Assume that his claimed occupational choices are business, engineering, and "executive work." He feels the need of help in making a final occupational choice.

At the point of interest test interpretation, the counselor can make this bald statement: "You have the interests of a Y.M.C.A. secretary or a personnel manager!" With minor modifications this is probably the standard approach to interpretation. There is no more probable way to lose a case than this. It is the least effective clinical approach, for the following reasons:

1. The student's spoken or unspoken response is usually "How can you say that? I never was a Y.M.C.A. secretary or a personnel manager!" At this point the counselor must backtrack and start a rather incoherent explanation of the basis of interest measurement, to his own and the student's confusion.
2. If the student accepts the statement without raising the foregoing issue in some form, the chances are he will re-interpret the statement, then or later, to mean that he has the *ability* to be a Y.M.C.A. secretary or a personnel manager, and that these are two jobs where his success is guaranteed. If any other factors interfere with curricular

*This article is the first draft of a chapter in a forthcoming monograph entitled: *Clinical Aspects and Interpretation of the Strong Vocational Interest Blanks*. Other theoretical and interpretive phases of interest measurement are treated more extensively in the monograph, to be published by the Psychological Corporation.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

or job success, he claims he "was told he would succeed in these jobs."

3. Such statements run the risk of flouting countless stereotypes, prejudices, specific dislikes, or misconceptions evoked by occupational labels, either in the student or in his parents. Very few people know what a personnel manager does, and there are substantial and not always complimentary stereotypes about Y.M.C.A. workers. The interesting fact that these labels are directly at variance with the student's claimed choices operates also to set up resistances in the student, although the student's own specific choices are also hedged around with favorable stereotypes which may be equally invalid, and although there is considerable evidence on the instability and invalidity of the student's claimed choices.

4. Such statements run the risk of moving the discussion too early in counseling to the temporarily irrelevant factors of opportunities, salaries, prestige values. The counselor is forced to waste precious time giving data (if he knows of any) on these points before having established an understanding of the interest type being discussed.

5. Such statements fail to take into account the vital factors of *levels* of ability and past achievement, which determine the level of future academic achievement most probably attainable; educational disabilities affecting educational progress in a correct curricular and occupational area; amounts of relevant specific aptitudes, *in addition* to level of general scholastic ability; and personality characteristics related to job success or satisfaction. Specific patterns of interest unaccompanied by ability and past achievement sufficient to permit curricular competition in professional schools occur frequently in counseling, because of the relatively low general correlations between measured interests and measured abilities or achievement.

Strong has published correlations of each occupational key with an intelligence test.¹ On the original blank the zero-order correlations range from $-.36$ to $.38$. Segel and Brintle² collected interest test scores, college grades and achievement test scores from 100 junior college freshmen. Using interest test scores for the keys for doctor, lawyer, life insurance salesman, personnel manager, and purchasing agent, they found only one positive correlation above $.40$ with selected parts of the Iowa High School Content Examination—the correlation between engineering interests

¹ See *Manual for Vocational Interest Blank for Men*, original and revised blanks. Palo Alto: Stanford University Press.

² David Segel and S. L. Brintle. "The Relation of Occupational Interest Scores as Measured by the Strong Interest Blank to Achievement Test Results and College Marks in Certain College Subject Groups." *Journal of Educational Research*, XXVII (February, 1934), 442-45.

COUNSELING ON BASIS OF INTEREST MEASUREMENT

and measured achievement in mathematics. Achievement in mathematics and science correlated .28 and .29, respectively with measured interests in medicine. Achievement in English literature, science, and social studies correlated —.43, —.26, and —.26 respectively with measured interests of a purchasing agent. The correlations between subject matter grades and measured interests were even lower than those between achievement tests and measured interests. Grades in mathematics and science correlated only to the extent of .14 with interests in engineering, while grades in history correlated —.47 with interests in engineering. The authors were sufficiently encouraged by these relations between scholastic accomplishment and interest test scores derived from studying adult occupational groups to suggest that "scales for scoring the Strong Interest Tests should be devised for the principal subject groups in higher secondary education." However, the obtained correlations were so low that the clinician must be extremely careful to keep interests and abilities or achievement separate in his own thinking, and to see that there is no such confusion in the student's thinking.

This error in counseling is particularly tragic and inexcusable where the occupations being discussed in terms of the interest test are those for which society demands college training prior to certification for professional competition. It is equally inexcusable in cases where the occupation can be entered with or without specific advanced training, as in the case of general measured interests in business. But in such cases, the counselor can cover his error by saying later what he should have explained earlier, namely, that in such occupations, success or satisfaction in the occupation is still possible even though success or satisfaction is not possible in a curriculum which may bear some degree of resemblance to the occupation, but which is not yet an indispensable prerequisite. This explanatory technique can be effectively used in "downgrading" some cases.

6. Such statements also fail to take into account the problem mentioned earlier³ in regard to the present-day representativeness of norm groups, as exemplified in the psychologists' key.

7. Finally, such blunt statements omit consideration of possible changes of specific measured interests which, while infrequent, may occur under certain conditions. Strong states this position clearly: "Prognostication of future behavior cannot safely be based upon the presence or absence of any single interest, but it does appear that to a consid-

³ The monograph from which this chapter is taken discusses the representativeness of Strong's standardizing groups as a factor in interpretation of the interest test results.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

erable degree at least it can be based upon the entire constellation of interests."⁴ In the article quoted, test-retest correlations of the specific keys ranged from .59 to .84 over a five-year interval beginning with the senior year in college.

Furthermore, in using the blank with younger students, it is usually more important to determine the interest type than the specific occupational interest. Carter, Pyles, and Bretnall⁵ have demonstrated the presence of the types at the average age of 16.5, whereas Carter and Jones⁶ have shown that only 17 per cent of tenth-grade students receive specific A scores on the keys appropriate to their occupational choices. Thus the counselor who uses the test with younger cases must remember that the standardizing procedure based on levels of scores made by adults may not yield an A score to a high-school student on a key within the interest type in which he may have a legitimate and dominant pattern.

With this understood, the test becomes clinically useful in the age range from about 15 years and up. But the counselor who looks only for single A scores cannot make effective use of the test in this age range. This difficulty would be clearly eliminated if a technique such as standard scores could be used as the reporting device for younger cases. Then the higher pattern of standard scores within an occupational group would stand out more clearly on the individual's profile, where the letter grade scores, based on adult norms, do not show intra-individual patterns so clearly in younger cases.

These statements of the ineffective way to interpret interest test scores, and the reasons therefore, grow out of bitter clinical experience. There is fortunately a more effective alternative. Suppose, in this hypothetical case, no reference is made to the interest test scores until *late* in the counseling interview. Suppose, further, that the counselor draws out of the student, by questioning, the reasons behind the student's own choices of business, engineering, or "executive work." He will discover much superficial thinking about jobs, which is in itself important. But he will also discover the specific factors leading to the choices: information (or misinformation) regarding salary scales and "over-

⁴ E. K. Strong, Jr., "Permanence of Vocational Interests," *Journal of Educational Psychology*, XXV (1934), 336-44.

⁵ H. D. Carter, G. K. Pyles and E. P. Bretnall, "A Comparative Study of Factors in Vocational Interest Scores of High-School Boys," *Journal of Educational Psychology*, XXVI (1935), 81-98.

⁶ H. D. Carter and Gary G. Jones, "Vocational Attitude Patterns in High-School Students," *Journal of Educational Psychology*, XXIX (1938), 321-35.

COUNSELING ON BASIS OF INTEREST MEASUREMENT

crowded" or "undercrowded" fields and job duties; satisfaction expected from the job; self-estimates of strong and weak abilities or subject-matter fields; evidences of family pressures or traditions dictating the choices; self-estimates of aspirations and motives that are operative in the choices; and evidences of out-of-school experiences shaping the choices.

Suppose, finally, that the counselor is familiar with the "interest types" or "interest patterns" growing out of factor analysis studies.⁷ The counselor can then direct the questioning at getting the student to evaluate activities which are related to the interest type and which are within the scope of his experience with his environment. Questions can also be used to evaluate those experiences contra-indicating the type into which the student's claimed choices fall.

Specifically, in the hypothetical case, unhappy experiences with mathematics would contra-indicate the technological interest type, in which the *claimed* choice of engineering is included. Participation in Hi-Y work and summer camp jobs may be drawn out as bits of evidence in favor of the welfare or uplift type in which some of the *measured* interests fall. A discussion of "executive work" as a pervasive problem of dealing with people takes it out of the *claimed* realm of a business activity alone.

Notice that the student has not yet been informed of his own specific measured interests. Notice also that the counselor has used the test scores in directing his questions to evoke relevant experiences and to clarify the student's thinking about jobs. At or near this point, the counselor will be ready to tell the student what his basic interest type seems to be, with some chance of getting this idea across by saying: "It seems to me that your basic interests are in helping people or in working with them

⁷Available factor analysis studies establish the qualitatively different types of interest patterns somewhat as follows: interest in scientific or technological activities; interest in verbal or linguistic activities; interest in business contact activities; interest in business detail activities; interest in welfare or uplift activities. The specific occupational keys for which the men's interest test is scored may be approximately grouped in these five categories. To make a clinical determination of the intensity of the interest type, the following procedure has been used with experimental verification, including tabulations of frequency of occurrence: for an individual student, the *primary interest pattern* is the interest type within which he shows a preponderance (majority or plurality) of A and B+ scores on the specific occupational keys; the *secondary interest pattern* is the interest type within which he shows a preponderance of B+ and B scores; and the *tertiary interest pattern* is the interest type within which he shows a preponderance of B and B— scores on the specific keys.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

in an effort to bring about an improved adjustment, rather than in technical, impersonal activities, or in piling up a tremendous fortune." Then he may discuss specific occupational duties and labels as representatives of the basic type, phrasing his remarks somewhat as follows: "These basic interests in helping people (or working with people) would be satisfied by the job of a personnel manager, for example, who is responsible for . . ." (and then may follow a description of job duties and responsibilities and types of training) . . . ; or those same interests would find an outlet in the type of work that a Y. secretary might do. So far as training is concerned, these two jobs require somewhat different types of abilities and aptitudes as we can see in studying the two curricula involved; therefore, it is important to see how your abilities and past achievements line up with the two choices "

In this way the A and B+ scores are introduced as examples of occupational outlets for the interest type rather than rigid occupational prescriptions for this student, and due allowance can be made for existing curricular differences.

The advantages of this clinical procedure are obvious. It reduces to a minimum the arousal of resistances growing out of stereotypes or prejudices which the student may have about the occupational label. It permits the counselor, subject to his own imagination and knowledge of jobs, to generalize beyond the available keys on the blank and classify other occupations within the basic interest type, which is valuable when one realizes that there are about 20,000 occupational labels and only 36 occupational keys on the revised blank for men, and 17 occupational keys on the blank for women. It permits the counselor then to discuss levels of ability, achievement, and aptitude required for a wider range of jobs within the interest type, and thus it permits readjustments of the student's plans in the light of other pertinent data about him. It gives the student a clearer understanding of the place of interests in making a vocational choice, because the counselor can explicate the student's responses to his earlier questions as they relate to an interest type theory. It reduces to a minimum any conflict between the student's specific choices and the counselor's alternative suggestions, since both the specific choices and the alternative suggestions are assigned to broader categories of interest types, where the student can more easily see his own status in regard to types of occupations.

The clinical effectiveness of this alternative plan of interpreta-

COUNSELING ON BASIS OF INTEREST MEASUREMENT

tion has been demonstrated in the experiences of graduate students in supervised clinical training, and in the reaction of trained counselors to the plan. Students are less prone to misinterpret the outcomes of the interview; parents can see more clearly the relevance of specific educational and vocational suggestions made by the counselors; greater flexibility is possible in working out educational and vocational plans; more satisfaction is expressed by students with this form of counseling assistance in their vocational problems.

No claim of infallibility, however, is made for the plan of interest test interpretation. It is not easy to learn, nor will it solve certain student problems of inflexible and over-emotionalized or fixated vocational choices. It requires skillful interviewing and careful explanations.

There are other aspects in counseling on the basis of interest measurement that should be mentioned. The absence of a consistently significant correlation between specific occupational scores and either ability or achievement has already been mentioned. Yet in these studies certain experimental problems remain uncontrolled. Clinicians can cite many cases in which a student has substantially improved his college grades when he transfers to a curriculum that trains for an occupation which is within his basic and primary interest type. Students transferring from engineering to business administration, from medicine to journalism, from chemistry to teaching, and succeeding better after such transfers are familiar to all counselors. The grade increment cannot be attributed solely to easier academic competition in the second curriculum, since the second curriculum may demand no less general academic ability than the first, and may demand different types of special achievements and aptitudes than the first.

If the interest measurement can be considered an approximate quantification of motivational factors, the following experiment would be significant. Choose a group of students having a primary pattern and a group having a secondary or tertiary pattern in the interest type for which a given curriculum offers specific occupational training. Match cases from the two groups on the basis of scholastic ability. If the primary pattern in the interest type denotes more adequate motivation, the group having this pattern should earn better grades than its matched group, provided no disproportionate factors of disabilities or problems load this experimental group.

Furthermore, when any raw score on an interest key above

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

—5 sigma can receive the same A grade, there is some question of the legitimate use of the Pearsonian correlation in studying the relation of occupational interest scores to other and more normally distributed variables, such as ability or achievement. Correlation ratio or contingency coefficient statistics may be more appropriate forms of analysis for these data. It is too early to consider occupational interest factors conclusively unrelated to curricular factors, in the light of examples to the contrary.

Conversely some counselors can cite cases in which superior or adequate grades are earned in a curriculum that trains for a vocation included in an interest group where the student has no primary pattern. Yet this need not be too alarming in the light of subsequent data about the occupational adjustments of graduates. Approximately fifty per cent of all the engineering graduates do not continue through life in the technical practice of engineering, and the chances are good that many in that fifty per cent have primary patterns in interest types other than the technical type.

This leads to a final clinical aspect. The interest type in counseling must be considered in relation to the local institution's curriculum organization in educational guidance. Examples are more clearly seen in terms of the blank for women. Many women make a primary pattern in the interest type which includes the secretarial and office worker keys. The normal curricular path in college may be the highly theoretical and technical economics of the existing school of commerce or business administration. Yet only a small proportion of college girls want to swallow this large dose of abstruse economic theory. The primary interest pattern is still a true picture of the occupational activities that would be satisfying prior to marriage; the curriculum may still be excellent for professional specialists in business, but the twain probably shall not meet happily. General education courses plus a minimum of training in basic office skills would solve the problem if the institution provides such a curricular organization. Otherwise a liberal arts education with a short course in a commercial business college during the summers or after school will provide a workable solution.

Other examples will occur to counselors who are familiar with the detailed structure of their curricula as well as the curricular labels. The general principle of empiric identification of interest type consonant with curricular duties will clarify some confusing cases for counselors.

THE COURSE IN SELF-APPRAISAL AND CAREERS OFFERED TO SENIORS IN THE CHICAGO PUBLIC HIGH SCHOOLS*

GRACE MUNSON

Bureau of Child Study, Chicago Public Schools

Since February, 1939, seniors in the Chicago high schools have been given the opportunity to enroll in a course in Self-Appraisal and Careers. This course, with its subsequent counseling, constitutes the final step in the Adjustment Service. It is the culmination of the self-appraisal and educational planning which starts early in the elementary grades, is featured in the eighth-grade program of articulation between elementary and high schools, is an important aspect of the individual counseling at all year levels by high-school teachers in their daily adjustment periods, and is featured again in the third-year program for a re-check on mental abilities and reading achievement. These activities are given continuity by the cumulative folder system and are supplemented all along the way by the individual service and follow-up studies of both elementary and high-school adjustment teachers collaborating with the Bureau of Child Study psychologists and demonstrators, for individual cases studies, clinical treatment, and consultative service.

As the Adjustment Service has now been operating in the high schools since 1937 and in the elementary schools since 1936, cumulative folders of the fourth-year students of September, 1940, will contain data assembled over a period of three and a half years and in some cases longer. Each year the data will extend back farther until ultimately the complete school history with many successive measures of mental power and achievement will be available for the final guidance step.

Given in the first half of the fourth year, the course in Self-Appraisal and Careers enhances and continues the self-appraisal of the earlier years by presenting the concepts of mental growth, of individual differences, and of the forces of self determination. It makes use of a wide range of scientific measuring techniques administered by the adjustment teacher or field psychologists for

* This article is a summary of the description of the course as presented in the Superintendent's Annual Report for the year 1939-40.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

the identification of specific areas of mental power, aptitudes, academic masteries, and vocational interests. And it teaches the techniques for interpreting the profiled results and the various conditioning factors. The self-appraisal section of the course gives each student a foundation in elementary psychology as a background for developing the techniques which will enable him to make continued self-appraisal as his pattern of powers and achievements changes with new growth and new experiences. The new understandings and newly accumulated data together with that assembled through the years are now used for making specific immediate plans and tentative future plans for educational, vocational and avocational pursuits.

Career Study Is Dynamic

The careers section of the course provides studies in specific vocational areas using the most recently published books and pamphlet series, compilations of current occupational information, regional conferences with selected speakers from representative vocational areas, personal interviews with these speakers, radio broadcasts, and tours. Students acquire knowledge of the historical development of occupations, their social significance, legislative controls and significant trends as a background for the development of techniques which will prepare them to continue the study of vocations on the basis of the new experiences and the new skills that may be acquired in the changing and diversified world of work.

The careers section of the course now lacks roots in earlier vocational studies comparable with the early development of self-appraisal. The problem of adjusting the high-school curriculum to accommodate such a course for all students earlier than the senior year has been studied with great care, since too early selection of vocations is detrimental, yet tentative choices should govern to some extent educational planning in high school. Beginnings have been made by introducing into selected subject-matter courses, study units on the vocational implications of a particular subject; books on vocations have been added to the free-reading book shelf for the first-year English Reading classes, and school libraries contain many valuable sources of vocational information; the individual counseling at the eighth-grade level and successive high-school levels by adjustment teachers, division-room teachers, and particularly by placement counselors involves some future vocational planning with the students; but

THE COURSE IN SELF-APPRAISAL AND CAREERS

the students, having had little opportunity to study careers, are unable as yet to contribute intelligently their rightful share of the planning.

The development of a program preliminary to the fourth-year course is necessary since an earlier tentative selection of a career plan will contribute to good mental hygiene by developing security, responsibility, organization of effort, and growth in self-determination. In this connection, plans are being formulated to drop the third-year testing program to the first half of the second year, using the New Chicago Tests of the Primary Mental Abilities which will yield more diversified data as a basis for self-appraisal and counseling; to introduce more specific career studies in the second-year curriculum together with a study of the total high-school organization of courses and facilities; and to establish more clearly defined routines to govern individual program-making from year to year by division-room counselors.

In the senior course the psychological studies and the careers studies are presented in somewhat parallel order, one vocational area being finally selected for intensive study by each student, after the results from the psychological measurements have been profiled and interpreted by him. A sample profile is presented in Figure I.

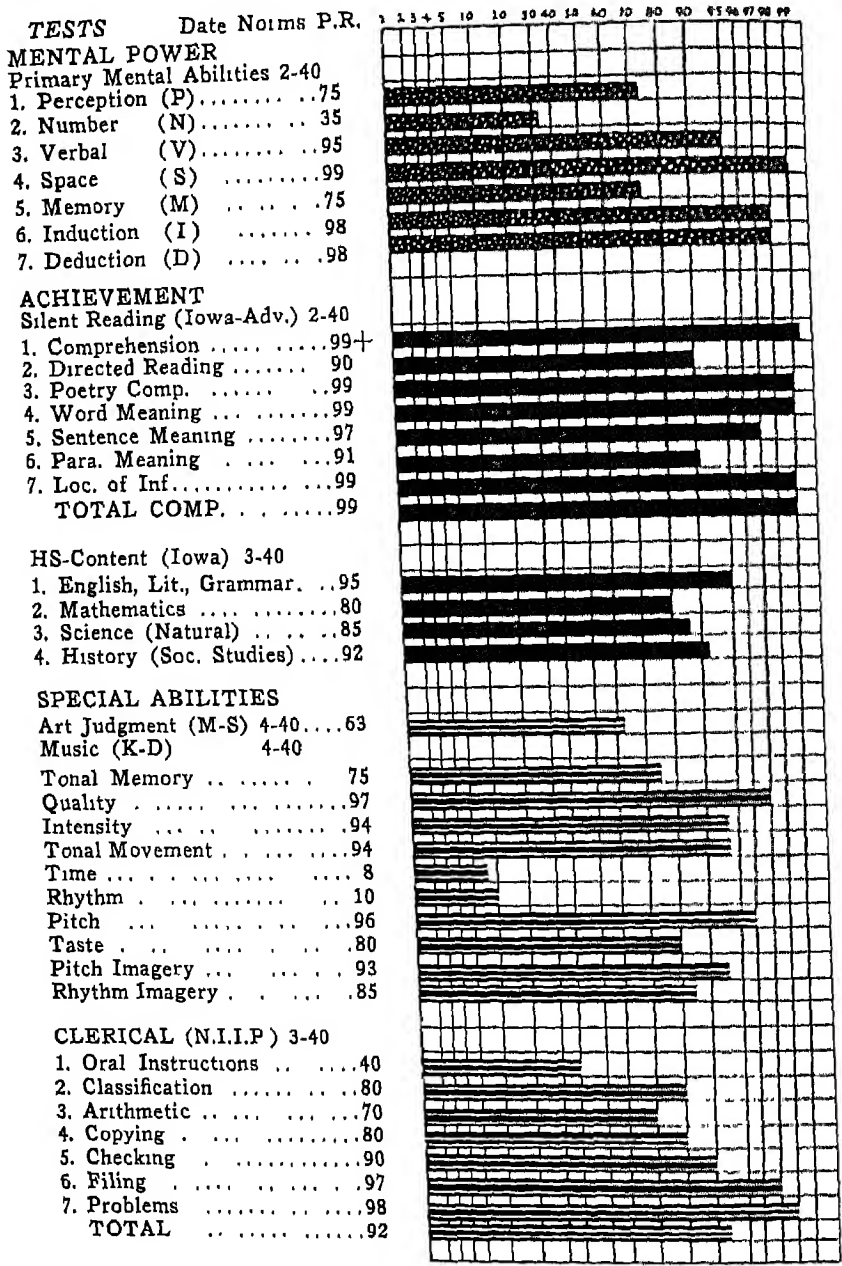
There is little attempt to match a given profile pattern to a particular vocation since scientific research has not been able to map the specific mental abilities required for insured success in a given vocation, and since an attempt to sort individual students into vocational pigeon holes would violate the fundamental democratic principles of public education. Yet wise counseling combines with student freedom of choice based on a knowledge of self and of careers, to give each student the security of tentative but specific plans.

The teachers of the course assist students in the formulation of such plans through individual counseling as the course progresses, using their non-teaching periods for this purpose. In the following semester each student goes over his plans again with the placement counselor if he seeks employment immediately following graduation, or with the senior counselor or the adjustment teacher, selecting his college and his first college courses, if he plans to continue his schooling. Most adjustment offices maintain a library of college information and scholarship data. The adjustment teachers arrange for senior visiting days at the junior colleges, and confer with their personnel staffs for

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Figure I

Profile of the test results for a student in self-appraisal and careers.
Name _____ School _____ Date _____ Period _____ Grade _____



THE COURSE IN SELF-APPRAISAL AND CAREERS

the orientation program and the transfer of data for such students as plan to attend. They also assist in organizing "College Day" when representatives from local and state colleges and universities present the advantages of their institutions.

The course in Self-Appraisal and Careers has been organized and serviced by the Bureau of Child Study and the Bureau of Occupational Research subject to the advice and guidance of the Assistant Superintendent in Charge of High Schools. Conferences with principals have determined policies for the outlines and content of the course while the teachers have contributed many valuable devices and suggestions. Each semester the students have made constructive criticism to improve the value of the course for the next group.

Since the course is a five-hour major elective it has not been accessible to all seniors following the old program of high-school studies. The new program, which will begin to operate in the next semester and which allows wider choice of electives, will permit more students to enroll. The course should eventually be made available to every fourth-year student.

The following table shows the enrollments in successive semesters since the course was established in February, 1939:

Enrollments in Self-Appraisal and Careers

Calendar	No. of Schools	No. of Classes	No. of Students
February, 1939.....	32	74	2600
September, 1939....	32	69	2500
February, 1940....	36	80	2800

The course is taught without a textbook since no high-school textbook has been written that covers the psychological studies selected for the course and since most textbooks on occupations are likely to be out of date by the time they are printed. Instead, an extensive bookshelf of reference materials for both teachers and pupils is supplied, supplemented by current materials on occupations. Students thus have an opportunity to read widely in the areas of their interests. To make the books more available to students, several books have been unitized for each school, by dividing them into from 14 to 43 sections re-mounted in manila covers, thus introducing a type of individualized instruction. This year a set of 10 reprints on psychological topics, written in popular vein by eminent leaders in that field, was supplied in class sets to each high school.

Teachers' lesson plans and outlines have been worked out and

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

distributed to all of the schools, modified from semester to semester in accordance with suggestions from teachers and pupils. This year, in answer to the demands, student work sheets were prepared to accompany the teachers' outlines. They were mimeographed and supplied in class sets.

The Outline of the Course

The outline of the course, arrived at by successive modifications in the light of experience, is given below. It will be revised still further as experience indicates the directions in which it may be more useful to the students and more adequate in fulfilling its objectives. Lesson outlines for the teacher have been prepared for all sections of the course. Special study guides have been prepared for the use of students in connection with the topics which are starred. Units I and II have been prepared by the Bureau of Child Study, Unit III by the Bureau of Occupational Research, Unit IV jointly.

Unit I. *Introduction and Bibliography

- A. Aims and activities of the course
- B. Terminology
- C Bibliography

Unit II. Self-Appraisal (To be taught simultaneously with Unit III. It is suggested that each week, two days be spent on Unit II, two on Unit III, and one on testing. Constant interweaving should be practiced.)

- A. Existence of individual differences
 - 1. Family history and autobiography
 - a. Racial and cultural background
 - b. Family traits, vocations and achievements
 - c. Health history
 - d. Educational history
 - e. Hobbies
 - f. Social development
 - g. Occupational experiences
 - h. Plans for the future
 - *2. Physical and mental differences between people
 - a. Types of differences
 - b. The total personality
 - c. The normal curve
 - d. Applications to the testing program
 - e. Educational implications
 - f. Vocational implications

*Special study guides have been prepared for the use of students in connection with the topics which are starred. Lesson outlines for the teacher have been prepared for all sections of the course.

THE COURSE IN SELF-APPRAISAL AND CAREERS

- g. Chicago's plan for the study of individual differences from the kindergarten through the high school
- ¹B. Uses and limitations of standardized tests
 - 1. How accurate are the test results?
 - 2. How useful are the test results for prediction?
 - 3. Do the tests measure all one's abilities?
 - 4. How can the information from them be used most effectively?
 - 5. Can the tests designate the one particular job for which each person is exactly fitted?
 - 6. Study of tests to be given in this course
 - a. Description
 - b. Why selected
 - 7. Chicago plan for the study of individual differences and for the development of techniques of self-appraisal from fourth grade through high school
- C. Psychological factors that must be considered in the interpretation of test results
 - 1. Maturation and change
 - *a. The process of growing up
 - (1) Physical and mental growth
 - (2) Laws of natural growth—infancy to maturity
 - (3) Influence of the environment on growth
 - (a) Effect of frustrations
 - (b) Effect of social environment
 - (4) Adolescence
 - (5) Maturity—the learning ability of adults
 - *b. Individual control of the direction of growth
 - (1) Habits: our masters or our servants
 - (a) Conditioned response
 - (b) Deliberate reconditioning
 - (2) Development of work habits
 - (a) Urge to mastery and completion
 - (b) Urge to self-direction
 - 2. Mastering our environment
 - *a. Human drives and obstacles
 - (1) The basic drives
 - (2) Psychological bases for the emotions
 - (3) Motives derived from basic drives
 - (4) The complexity of motives
 - (5) Motives as products of the environment, plus psychological factors
 - (6) The universality of obstacles
 - (7) Drives in career planning
 - *b. Mastery and adjustment—interaction of the individual and the environment

¹ See footnote on preceding page.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

- (1) Successful adjustment: mastery of one's problems
 - (a) Adjustment of environment to self
 - (b) Adjustment of self to environment
 - (2) Less successful types of adjustment
 - (3) Self-appraisal and the choice of adjustment
 - (4) Relation to educational and vocational planning
- D. Individual interpretation of the test results
1. Necessary statistical concepts
 - *a. Percentile rank
 - b. Mean
 - c. Median
 - d. Quartile
 2. Profiles of test results to be made by the student
 - a. Construction
 - b. Interpretation of test data as samplings
 - c. Comparison of abilities and achievements
 3. Aids in interpretation of individual performance on each of the following tests:
 - *a. Thurstone Primary Mental Abilities
 - *b. American Council on Education Psychology Examination
 - *c. Iowa Silent Reading Test, Advanced
 - *d. National Institute of Industrial Psychological Clerical Examination
 - *e. Cleeton Vocational Interest Inventory
 4. Aids in interpretation of the completed profile
- Unit III. Careers and Occupations (Simultaneously with Unit II)
- A. Man's interdependence in work
1. The growth of interdependence
 - a. Primitive methods of work
 - b. Development of specialization
 - c. Effect of specialization
 - d. Discussion of our present highly specialized working world
 - e. Release of human energy for cultural service, and leisure-time activities
 2. Evolution and importance of occupational groupings
 - a. Development and significance of the merchant guilds
 - b. Development of craft guilds
 - c. Later history of craft guilds
 - d. Present day significance
 - (1) Employee organizations
 - (2) Employer organizations
 - (3) Trade associations
 - (4) Professional organizations

* See footnote on page 48.

THE COURSE IN SELF-APPRAISAL AND CAREERS

3. Socio-economic factors in the study of an occupational area
 - a. Questionnaire study
 - b. Basic attitudes toward occupational rewards other than money
4. Legislation affecting workers
 - a. Social Security—old age insurance
 - b. Social Security—unemployment compensation
 - c. Wage and hours laws
 - d. Child labor laws
- B. Significant relations and trends in occupations
 1. Classification of occupations
 - a. Importance of study of occupational areas in a broad sense as well as of specific occupations
 - b. Occupational areas vs. occupational fields
 2. Significance of trends in occupations
 - a. Technological
 - b. Commercial
 - c. Personal and domestic
 - d. Professional and semi-professional
- C. Study of an occupation
 1. Relationship between the school subjects and occupations related to those subjects
 2. Graphs of life earnings
 3. Case study of an individual
 4. Intensive study of several selected occupations (check list or outline for occupational study)
 5. Intensive study of several selected avocations
- D. Techniques in securing and holding work
 1. Channels in finding work
 2. Written application for work
 3. Making an interview
 4. Adjusting to a job

Unit IV. *Summary

- A. Development of techniques for self-guidance
 1. Summarizing self-appraisal data
 2. Summarizing data for study of occupations
- B. Schedules for counseling during the ensuing semester
 1. Functions of placement counselor
 2. Functions of other counselors available in school
 3. Appointments
- C. Application of data to the solution of individual problems and the formulation of two plans—one a tentative long-range plan, and one a specific plan for immediate action, both to include provisions for education, vocation, and avocation

* See footnote on page 48.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

D. Evaluation of the course

Student Work Sheets have been prepared to implement the outline of Units I, II, and IV. The topics are listed below. Over 2,000 copies of each were distributed to students during the year 1939-1940.

1. Introduction and Bibliography, 17 pp.
2. Physical and Mental Differences, 6 pp.
3. Standardized Tests, 7 pp.
4. Process of Growing-Up, 15 pp.
5. Self-Directed Personality Change, 10 pp.
6. Human Drives, 14 pp.
7. Mastery and Adjustment, 10 pp.
8. Meaning of Percentile Rank
9. Primary Mental Abilities, 5 pp.
10. Mental Power as measured by A.C.E. Test, 4 pp.
11. Reading Ability as measured by the Iowa Silent Reading Test, Advanced, 5 pp.
12. Clerical Ability as measured by the N.I.I.P. Test, 4 pp.
13. Vocational Interest, as indicated by the Cleetson Vocational Interest Inventory, 7 pp.
14. Summarizing Self-Appraisal Data, 2 pp.

Battery of Tests for Self-Appraisal Used in 1939-1940

I. Mental Tests

A. *Thurstone Tests for the Primary Mental Abilities

1. Perception
2. Memory
3. Number
4. Space
5. Verbal
6. Inductive reasoning
7. Deductive reasoning

B. *American Council on Education Psychology Examination

II. Reading Ability

A. *Iowa Silent Reading Test, Advanced

III. Achievement Tests

A. *Iowa High School Content Examination

B. American Council on Education Cooperative General Achievement Test

1. Mathematics
2. Science
3. Social Science

IV. Aptitude Tests, as desired

A. Clerical

*National Institute of Industrial Psychology Clerical Test, American Revision

*Percentile norms have been prepared by the Bureau of Child Study on Chicago groups.

THE COURSE IN SELF-APPRAISAL AND CAREERS

- B. Mechanical
 - 1. Detroit Mechanical Aptitudes Examination
 - 2. Individual Manipulation Tests
- C. Musical
 - 1. Kwalwasser-Dykema Music Tests
 - 2. Seashore Measures of Musical Talent
- D. Artistic
 - Meier-Seashore Test for Art Judgment
- V. Miscellaneous
 - A. Cleeton Vocational Interest Inventory
 - B. Business Education Council Personality Rating Schedule
 - C. Kuder Preference Record

Bookshelves for Students and Teachers

The following bookshelves have been set up for students and teachers: One set has been furnished to each high school.

References for Students

Psychological Books

- *Blatz, William E. *The Five Sisters*. New York: W. Morrow and Company, 1939.

Psychological Pamphlets (30 sets to each school)

- *(Reprints from a series of radio lectures published under the title of *Psychology Today* by the University of Chicago Press, 1932.)

- Garrett, Henry E. *Psychology Today*
Goodenough, Florence *Child Development*
Gesell, Arnold *Growth of the Infant Mind*
Watson, John B. *How to Grow a Personality*
Allport, Floyd H. *Personality in Our Changing Society*
Cannon, Walter B. *Effects of Strong Emotion*
Warden, Carl J. *Animal Drives*
Robinson, Edward S. *Learning and Forgetting*
Thorndike, Edward L. *Effects of Rewards and Punishments*
O'Rourke, L. J. *Matching Men and Occupations*

Occupational Books

- †Brewer, John M. *Occupations*. Boston: Ginn and Company, 1937.
†Chapman, Paul W. *Occupational Guidance*. Atlanta: Turner E. Smith and Company, 1937.
†Clark, Harold F. *Life Earnings*. New York: Harper and Brothers, 1937.
†Fleischman, D. E. *An Outline of Careers for Women*. Garden City: Doubleday, Doran and Company, 1935.
Giles, I. K. *Occupational Civics*. New York: Macmillan Company, 1936.
*Lyons, George J. and Martin, Harmon C. *The Strategy of Job Finding*. New York: Prentice-Hall, Inc., 1939.

* Books added during 1939-1940.

† Books which have been unitized.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

National Resources Committee *Technological Trends and National Policy*. Washington, D. C.: United States Supt. of Documents, 1937.
Occupational Outlines on America's Major Occupations. Chicago: Science Research Associates, 1940.

†Rosengarten, W. *Choosing Your Life Work*. New York: McGraw-Hill, Inc., 1936.

United States Dept. of Commerce *Census of Business: 1935*. Washington, D. C.: Bureau of the Census, January, 1937. Also *Census of Retail Trade, 1936*.

†Williamson, E. G. *Students and Occupations*. New York: Henry Holt Company, 1937.

†Ziegler, S. H. and Wildes, Helen J. *Choosing an Occupation*. Philadelphia: John C. Winston Co., 1937, revised edition.

Occupational Pamphlets

American Job Series. Chicago: Science Research Associates, 1700 Prairie Avenue. 19 occupational monographs.

Are There Opportunities for Women? 1936. 10 pamphlets. *Changing Patterns in Occupations*. 1936. 26 pamphlets. New York: National Federation of Business and Professional Women's Clubs, 1819 Broadway.

Occupational Pamphlets. New York: National Occupational Conference. A series of appraisals and abstracts of available literature. 57 pamphlets.

Occupational Research Reports. Chicago: National Youth Administration of Illinois, Merchandise Mart. 29 pamphlets.

Occupational Briefs. Briefs compiled by the National Youth Administration on the occupations included in the reports above.

Guidance Leaflets. Washington, D. C.: United States Printing Office, 1936. 19 pamphlets.

Success—Vocational information series. Directed by Chloris Shade, Joliet Township High School. Chicago: Morgan-Dillon and Company. 55 pamphlets.

Bibliographical Helps

Bennett, Wilma. *Occupational and Vocational Guidance—A Source List of Pamphlet Material*. New York: H. W. Wilson Company, 1936, revised edition.

*Massachusetts Youth Administration, *Bibliography of Occupational and Apprenticeship Information*. Boston: 31 St. James Avenue, 1937. 101 pp. Comprehensive list of magazine articles.

Parker, Willard B. *Books About Jobs*. Published for the National Occupational Conference by the American Library Association, Chicago, 1936.

Price, Willodeen and Ticen, Zelma E. *Index to Vocations*. New York: H. W. Wilson Company, 1936, revised edition.

Bibliography of References on Vocational Guidance for Girls and Women. United States Office of Education. Washington: Vocational Division, 1936, revised, 13 pp. Lists bibliographies, studies and investigations.

* Books added during 1939-1940.

† Books which have been unitized.

THE COURSE IN SELF-APPRAISAL AND CAREERS

Vocational Guide. Chicago: Science Research Associates. A monthly bibliography of occupational books and articles.

Research Services

Occupational Card File on Current Local Data Chicago: Bureau of Occupational Research, Board of Education.

Cumulative Bulletin Series. Chicago: Bureau of Occupational Research, Board of Education.

Special Research Reports. Chicago: Placement Clearance Center, a division of the Bureau of Occupational Research, Board of Education.

References for Teachers

Psychological Books

Bingham, Walter V. *Aptitudes and Aptitude Testing.* New York: Harper and Brothers, 1937.

Paterson, Donald O., Schneider, Gwendolen G., and Williamson, E. G. *Student Guidance Techniques.* New York: McGraw-Hill, Inc. 1938.

Shaffer, Lawrence F. *The Psychology of Adjustment* Boston: Houghton, Mifflin Company, 1936.

Strang, Ruth M. *Role of Teacher in Personnel Work.* New York: Teachers College, Columbia University, 1936.

Occupational Books

Lincoln, Mildred E. *A Short List of References on Methods of Teaching Occupations.* New York: National Occupational Conference. Mimeographed, 3 pp. Free upon request.

Lincoln, Mildred E. and Brewer, John M. *How to Teach Occupations.* Boston: Ginn and Company, 1937.

The Reactions of Students

It is too early to obtain an adequate evaluation of the course. If the opinion of the students is a criterion (and who doubts that it is an important component?) the course is highly successful. The reactions of students indicate their deep sense of responsibility at this level of high school training, the changes in their viewpoints engendered by the course, and their gratitude both for the new knowledge acquired, and for the personal guidance from the fine men and women who have taught the course. A few of the student comments are presented below:

"I feel I have benefited by almost every topic and discussion in this course in Careers, but several parts have been very helpful. I enjoyed all the tests, and the experience of having had them helped me when I applied for a position at the Continental Bank. Four tests were required and one was almost identical to those we have been taking. All the tests at the bank were somewhat like those we have been taking and I was much more confident than I would have been, if the work had been new to me. Another part of the course that I feel has been of great help to me has been hearing

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

about various occupations. It is much easier to decide upon a vocation for yourself, one that you think you will like, after you hear the good and bad points about many vocations. I have a much clearer idea of what I would like to be in the future than I had before I took Careers."

Gloria K., Hirsch High School

"The survey of different vocations and professions has been most helpful to me. It never occurred to me that one vocation could branch into so many specific fields. The Careers class threw light on many subjects concerning which I was in the dark, such as the present and future demands in the labor market, the demand of the employer upon his employee, and the amount of education needed to get along in a vocation."

Frances D., Harper High School

"This course of Self-Appraisal which you are offering is very good in building character, citizens, and regular ladies and gentlemen. Now I don't say this just to be your good friend because it is everything that I mentioned above and more.

"One of the things that struck me most was the way you treat the pupils. Because there is nothing like having a regular guy talking to another regular guy.

"After three and a half years of bumming, cutting, etc., this course brought me to my senses. I don't know what it was — whether it was the tests, or the homely philosophy — but the course was interesting.

"And now in ending I want to thank you for making this change in me. And later in life I'll come up and give you a visit. Maybe I'll be a bum on Madison Street or a big shot on Michigan Boulevard. I'll always come and visit the regular guy and at the same time ask him for advice."

Ted T., Harrison High School

"Very few of those who finish high school ever sit down and take an inventory of themselves. The talks students have with the counseling teacher make you gather your wits about you and make you think of how to approach your employer. Those who are backward and bashful come out of their shell, due mostly to the reassurance of the teacher who gives them a boost upward."

Margaret I., Marshall High School

"The most important part of the Careers course is the making of the Career book on the selected occupations, because it helps you find out all about the occupation and to make sure you will fit in that line of work. After making my book on careers in pattern making, I found I needed to brush up on a few technical things. Some students found out that

THE COURSE IN SELF-APPRAISAL AND CAREERS

they never will or could fit in the occupation they had first chosen, and they have had time to make a better choice."

Chester L., Steinmetz High School

"Of all the courses I have had, one of the most important subjects for me and for the development of my character, has been "Self-Appraisal and Careers." It is a subject which we have to give a lot of thought to, with quite a bit of brain work. It has been helpful in many ways: (1) it makes one think fully on the future when he or she leaves education and goes into the open world; (2) it helps one to know people and to understand them much better; (3) it helps one discover the vocation into which he will best fit; (4) it helps one get a clearer view of the world and of occupations."

Kathleen M., Waller High School

Handbooks and Bulletins

More complete information concerning the course in Self-Appraisal and Careers will be found in mimeographed bulletins available from the Board of Education in the City of Chicago. The bulletins may be obtained for the cost of mailing (75c) by writing to the Bureau of Child Study, 228 N. La Salle St.

Prepared jointly by the Bureau of Child Study and the Bureau of Occupational Research:

Handbook on Self-Appraisal and Careers, 17 pp.

Teachers' Outlines for Self-Appraisal and Careers, 86 pp.

Prepared by the Bureau of Child Study:

Student Work Sheets for Self-Appraisal and Careers, 106 pp.

Handbook of Norms, 30 pp.

Handbook on Scoring Procedures, 36 pp.

High-School Teacher's Devices and Suggestions (Subject: Self-Appraisal and Careers. A Bulletin issued by the Superintendent of Schools), 19 pp.

Bureau of Child Study Annual Report, Part V, High-School Self-Appraisal and Careers Course, June, 1939, 9 pp.

Service Bulletin No. 1, 1940, Methods of Presenting the Course in Self-Appraisal and Careers.

Prepared by the Bureau of Occupational Research:

Cumulative Bulletin Series:

Series I Educational Facilities, 30 pp.

Series II Occupational Information, 23 pp.

Series III Significant Trends, 3 pp.

Series IV Pertinent Legislation, 11 pp.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

REFERENCES

- Johnson, William H, *Annual Report of the Superintendent of Schools*. Chicago: 1937-38 P. 339.
- . *Annual Report of the Superintendent of Schools*. Chicago: 1938-39. Pp. 176-79
- . *Annual Report of the Superintendent of Schools*. Chicago: 1939-40.
- . "Adjustment Service in the Chicago High Schools," *Educational Administration and Supervision*, XXIII (Oct. 1937), 513-20.
- . "Adjustment Service in High Schools," *American School Board Journal*, XCVI (May, 1938), 30-32.
- . "Adjustment Teacher Service in Chicago Elementary Schools," *The Elementary School Journal*, XXXVIII (Dec. 1937) 264-71.
- . "Guidance, Counseling and Adjustment," *The School Executive*, LVI (May, 1937), 333-34.
- . "New Self-Appraisal and Career Study Course in the Chicago High Schools," *School and Society*, XLIX (May 20, 1939), 627-31.
- . "Place of Guidance, Counseling and Adjustment in the Secondary Schools," *The North Central Association Quarterly*, XII (Jan., 1938), 369-72.
- Munson, Grace, "Adjustment Service of Chicago High Schools," *Occupations*, XVII (Feb., 1939), 389-94.
- . "Adjustment Service—Chicago Schools," *Educational Method*, XIX (March, 1940), 327-35.
- Schloerb, Lester J., "N.Y.A. Occupational Monographs," *Chicago School Journal*, XXI (Jan.-Feb., 1940), 180-81.
- . "Placement Counselors in Chicago Schools," *Occupations*, XVIII (Feb., 1940), 387.

PRIMARY MENTAL ABILITIES AND AVIATION MAINTENANCE COURSES*

WILLARD HARRELL, *University of Illinois*
and

RICHARD FAUBION, *Air Corps Technical Schools*

This investigation is the third of a series designed to determine the optimal pattern of abilities for mechanical work. The first study, "A Factor Analysis of Mechanical Ability Tests" (1) suggested that the principal component of the Minnesota series of mechanical tests is the Space factor. A second factor, tentatively identified as the Perceptual, was present in that battery. A Manual Agility factor was also isolated. None of the Minnesota tests possessed a significant weight for this Agility factor. The most practical conclusion from this first study was that certain paper and pencil tests will measure equally validly each of the factors present in more clumsily-administered mechanical tests.

The second study, "Selection Tests for Aviation Mechanics (2)," consequently involved only paper and pencil tests. This second study was started after the publication of Thurstone's monograph, "Primary Mental Abilities (3)," but was begun before his Experimental Battery of Primary Mental Ability Tests (4) became available. Nine of the tests from the monograph supplement were included along with 29 other sub-tests. These were taken by 84 basic instruction students of the Air Corps Technical Schools. Basic instruction grades from each of five aviation maintenance courses with a total duration of eight weeks formed external criteria. These course grades were the criteria for both the second study and for the third, the subject of this paper.

Air Corps Technical School students take these five basic instruction courses regardless of later specialization in radio, photography, airplane mechanics, parachute rigging or other advanced specialties. The five basic courses are Shop Mathematics, Mechanical Drafting and Blueprint Reading, Air Corps Fundamentals, Elements of Metalwork, and Elements of Electricity. The names are perhaps sufficiently definitive except for two of

* This report is of a study sponsored jointly by the Trade Test Department, Air Corps Technical Schools, and the University of Illinois' Graduate Research Committee. The paper was read at the Mid-Western Psychological Association, May 4, 1940, at the University of Chicago.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

the courses. Air Corps Fundamentals is unlike the others in that it does not entail mechanical problems. It is made up of the study of Air Corps rules and nomenclature. Shop Mathematics includes the following topics: addition, subtraction, multiplication, and division; fractions and decimals; denominate numbers and mensuration; formulas and tables; shop trigonometry; applied problems.

Entrance to the Air Corps Technical Schools is restricted to soldiers in the United States Army. Consequently the minimum age is 18 years. The education requirement is graduation from high school or the equivalent. A minimum Army Alpha percentile rank of 75 is required. Percentiles here are based on the Army population. The percentile rank of 75 corresponds to an Otis I.Q. of 100.

One hundred and five soldiers, students of the Air Corps Technical Schools, were given Thurstone's Experimental Battery of Primary Mental Ability Tests (4), and two additional tests found predictive in the previous study (2) — Surface Development and Punched Holes. Army Alpha scores were also available since they are used as an entrance requirement. About half the group was in the advanced phase of Airplane Mechanics, and the other was in Radio Mechanics. The age range was 18-39 with a mode of 19. The range for years of formal schooling was 9-15. Sixty-four had completed high school but had gone no further.

The classification of students into the various advanced phases, as well as their selection, might be considered a test problem in part, but only the selection angle will be considered here. Results are becoming available from tests given to 600 students to provide sufficiently large samples to trace the correlation between tests and several advanced phases.

It is recognized that the course grades are not perfect criteria. They are complex, but since they consistently correlate significantly with several tests, they probably possess a reasonable amount of validity. One objective criterion—a machine shop product—has been developed which is hoped to have a satisfactory reliability. Other practical criteria and objective information criteria are planned.

The reliability has been estimated by the split-half method for each of the sub-tests correlating as high as .30 with a criterion. These coefficients are shown in Table III.

Only two of the seven Alpha sub-tests with significant correlations with any grade, namely Addition and Analogies, have a

PRIMARY MENTAL ABILITIES AND AVIATION

reliability above .90. Alpha Arithmetic with reliability of .60 is lowest.

Three of the Primary Mental Ability tests, Completion, Arithmetic, and Number Series have reliabilities of less than .90 but more than .80. From an item analysis, showing the correlation of each item in the PMA battery with total sub-test score, the relatively low reliability in these three cases is probably due in part to the items not being arranged in order of their difficulty.

Four of the PMA tests, Addition, Same-Opposites, Cards, and Figures, have reliabilities above .97. These high reliabilities may be partially explained by the items within each of the tests being practically of equal difficulty.

Comparison of A. C. T. S. Students with High School Seniors

A comparison has been made between the PMA scores for Air Corps Technical School students and the norms published for 300 Hyde Park High School (Chicago) seniors. Table IV shows this comparison. Critical ratios have been calculated from the differences between means. CR's for Number and Memory are less than .30. Hyde Park seniors have higher Perceptual, Verbal, and Induction scores. Air Corps Technical School students have higher Reasoning and Space scores.

It is difficult to interpret these results because it is not possible to say exactly what selective agents are at work in the Air Corps Technical School. The most obvious ones are, being a soldier, choice by a commandant which presumably means interest in mechanical work, completion of high school, and having an Alpha Army percentile rank of 75.

A difficulty with the Reasoning or D score is that one of the tests, Mechanical Movements, on which the D score depends, also possesses a significant weight in another factor which from an unpublished factor analysis by the writers seems to be Knowledge of Mechanical Processes. Since the present group is selected in part on their interest and, presumably, knowledge of mechanical processes, this would increase the Mechanical Movements score and consequently the D score, without demonstrating that they are better reasoners than the Hyde Park seniors.

Results with Primary Mental Abilities Tests

All of the PMA scores were obtained from adding test scores. Five of the seven, all but Perception and Memory, correlate sig-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

nificantly with at least one of four basic instruction grades Table I lists the product-moment correlation coefficients A significant correlation is considered to be one where the coefficient is at least four times its probable error For 105 cases this is a correlation of .24

Elements of Metalwork does not correlate significantly with any of the tests, but it did in the second study referred to above. The test correlations with Shop Mathematics and with Mechanical Drafting appear quite similar to those in the previous study. In both groups, Addition, Number Series, and Surface Development correlated significantly with Shop Mathematics; and in each study Mechanical Movements, Surface Development, and Punched Holes with Mechanical Drafting. There are stronger correlations with Electricity, and with Air Corps Fundamentals in this study than in the second A possible explanation is that the present battery has more Verbal tests, and these correlate significantly with each of those two courses. More important is that the greater dispersion for age and schooling of the present group tends to increase the correlations.

Looking again at Table I, the Number factor correlates significantly only with Shop Mathematics; Space correlates significantly with Shop Mathematics, and with Mechanical Drafting; Induction with Shop Mathematics, Electricity, and Mechanical Drafting; while Reasoning and the Verbal score correlate significantly with each of the four basic grades

Multiple correlation coefficients using only significant zero-order coefficients have been computed between PMA scores and each of four basic grades These may be compared with correlations of Alpha total with the four criteria grades. The multiple R's from the factor scores are .46 with Shop Mathematics, .57 with Electricity, .60 with Mechanical Drafting; and .36 with Air Corps Fundamentals Corresponding values for Alpha total are .31, .47, .30, and .41 These multiple R's, as well as others to be mentioned later, would be expected to be less in other samples by the shrinkage effect if the same regression formulas were used

The multiple correlation between four factor scores, Verbal, Space, Induction, and Reasoning, is .63 with a composite basic grade obtained from adding grades in Shop Mathematics, Electricity, and Mechanical Drafting. Alpha total correlates .45 with this same composite The zero order correlations with this composite grade are given in Table VI. Table V shows the inter-correlations of five factor scores.

PRIMARY MENTAL ABILITIES AND AVIATION

Results are shown in Table II for those sub-tests which correlate 30 or more with one of the basic grades. This is five times the probable error.

Conclusions

The Air Corps Technical Schools are planning to supplement their test selection in line with these results; and they also expect to establish test standards for classification from future studies.

We have come to the conclusion from this and other studies that there is no one separate factor for a mechanical ability. Rather, there are several factors which are more or less prominent in mechanical work, their pattern depending on its type and complexity and on the point reached in the learning curve.

A Perceptual factor, although present in several so-called Mechanical Aptitude tests, is probably related to mechanical work, borrowing an expression from Holzinger, as an *Arti-factor*. The Verbal factor has been shown to be evident in training for mechanical work of relatively great complexity. Among the more important factors in mechanical operations are Space, one or two Reasoning factors, and Knowledge of Mechanical Processes. A Manual Agility factor is present in routine jobs where individual differences depend on the manipulation of objects such as nuts and bolts.

TABLE I

Product-Moment Correlation Coefficients Between 5 "Primary Mental Ability" Scores and 4 Aviation Maintenance Courses for 105 Soldiers*

	Shop Math	Elec- tricity	Blue Print Reading and Mech Draftg	Air Corps Funda- mentals
14 N (Addition, Multiplication)	37	17	00	11
17 V (Completion, Same Opposites)	28	51	37	33
20 S (Cards, Figures)	25	17	36	02
27 I (Letter Grouping, Marks, Number Patterns)	33	29	41	20
31 D (Arithmetic, Number Series, Mechanical Movements)	26	40	54	24
PE _r = .05 where r = .50				
PE _r = .06 where r = .20				

* Decimal points have been omitted before each coefficient.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE II

Product-Moment Correlation Coefficients Between 19 Tests and 4
Aviation Maintenance Courses for 105 Soldiers*

	Shop Math	Elec tricity	Blue Print Reading and Mech Draftg	Air Corps Funda mentals
1 Alpha Addition	33	06	- 10	-04
2 Alpha Arithmetic	21	35	30	27
3 Alpha Common Sense	13	31	06	06
4 Alpha Word Opposites	13	43	24	38
5 Alpha Mixed Sentences	17	43	23	34
6 Alpha Number Series	37	15	24	07
7 Alpha Analogies	26	35	23	39
12 Thurstone Addition	39	23	08	24
15 Thurstone Completion	23	47	39	21
16 Thurstone Same Opposites	27	45	30	34
18 Thurstone Cards	23	18	32	00
19 Thurstone Figures	21	12	33	05
24 Thurstone Letter Grouping	23	25	31	22
26 Thurstone Number Patterns	30	18	32	05
28 Thurstone Arithmetic	31	49	49	29
29 Thurstone Number Series	22	33	39	17
30 Thurstone Mechanical Movements	08	12	40	10
32 Thurstone Punched Holes	15	16	41	06
33 Thurstone Surface Development	35	21	50	02

* Decimal points have been omitted before each coefficient.

TABLE III

Test Reliabilities by the Split-Half Method (Stepped-up)
N = 103

1 Alpha Addition	98	19 Thurstone Figures	99
2 Alpha Arithmetic	60	24 Thurstone Letter Grouping	.91
3 Alpha Common Sense	87	26 Thurstone Number	
4 Alpha Word Opposites	88	Patterns	92
5 Alpha Mixed Sentences	80	28 Thurstone Arithmetic	.87
6 Alpha Number Series	84	29 Thurstone Number Series	.87
7 Alpha Analogies	93	30 Thurstone Mechanical	
12 Thurstone Addition	98	Movements	93
15 Thurstone Completion	.80	32 Thurstone Punched Holes	89
16 Thurstone Same-Opposites	99	33 Thurstone Surface	
18 Thurstone Cards	99	Development	.95

PRIMARY MENTAL ABILITIES AND AVIATION

TABLE IV

Comparison Between 300 Hyde Park High School Seniors and 105 Air Corps Technical School Students in "Primary Mental Abilities"

	Hyde Park Seniors		A C T S Students		CR
	Mean	S D	Mean	S D	
Perception	152	23.75	137.09	16.51	7.04
Number	119.5	30.00	118.78	27.80	0.22
Verbal	84.5	27.50	75.16	19.95	3.72
Space	109.5	35.00	125.23	34.14	4.04+
Memory	15.5	7.25	15.65	7.67	0.18+
Induction	35.5	9.00	28.46	8.76	7.04
Reasoning	54.5	18.75	68.85	18.48	6.69+

TABLE V

Product-Moment Correlation Coefficients
Among Factor Scores for 105 Soldiers

	N	V	S	I
V	31			
S	30	15		
I	33	28	39	
D	20	33	41	54

TABLE VI

Product-Moment Correlation Coefficients Between Tests and a Composite
Basic Grade Composed of Shop Math, Electricity, and
Mechanical Drafting

N = 105

1 Alpha Addition	.11	19 Thurstone Figures	.29
2 Alpha Arithmetic	.34	20 Space	.34
3 Alpha Common Sense	.17	24 Thurstone Letter Grouping	.34
4 Alpha Opposites	.31	26 Thurstone Number	
5 Alpha Mixed Sentences	.32	Patterns	.36
6 Alpha Number Series	.35	27 Induction	.44
7 Alpha Analogies	.35	28 Thurstone Arithmetic	.53
Alpha Total	.45	29 Thurstone Number Series	.39
12 Thurstone Addition	.29	30 Thurstone Mechanical	
14 Number	.23	Movements	.26
15 Thurstone Completion	.44	31 Deduction	.50
16 Thurstone Same-Opposites	.41	32 Thurstone Punched Holes	.31
17 Verbal	.47	33 Thurstone Surface	
18 Thurstone Cards	.32	Development	.47

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

REFERENCES

1. Harrell, T W "A factor analysis of mechanical ability tests" *Psychometrika*, V, 17-33
2. Harrell, T W and Faubion, R W "Selection tests for aviation mechanics" *Journal of Consulting Psychology*, in press
3. Thurstone, L L "Primary mental abilities" *Psychometric Monographs*, No 1 Chicago University of Chicago Press, 1938
4. Thurstone, L L *Manual of instructions Tests for primary mental abilities* Washington American Council on Education, 1938

A COMPARISON OF THE ORIGINAL AND REVISED STANFORD BINET INTELLIGENCE SCALES

MARTIN L REYMERT AND RALPH K MEISTER
The Mooseheart Laboratory for Child Research

The present study is an attempt to compare the original and the revised Stanford-Binet Intelligence Scales. The data have been obtained from 440 Mooseheart children, each of whom has had from two to nine examinations. The population of tests comprises 958 administrations of the original scale and 823 administrations of the revised. The testing was done by trained clinical psychologists of the Laboratory staff. The children are all normal and have been drawn from every state in the Union, predominantly from the Middle West.

The following items were recorded. The child's name, birth-date, the date of administration of the test, the I Q rating obtained, the M A, the C.A., the basal year score, the highest level of success and the amount of scatter. The time interval between administrations and the direction and amount of deviation from the first I Q rating to the second were obtained for each pair of successive administrations.

To compare the equivalence of ratings from one scale to the other with the respective reliabilities of the scales, using the same population, two groups of children were chosen. All had had at least two examinations with the original scale and two with the revised. However, Group A of Table I had taken the L form of the revised scales first while Group B had taken the M form first.

TABLE I
Correlations Between the Various Forms of the Stanford-Binet Scales
for Constant Populations

Scales Correlated	Group A			Group B		
	O ₁ O ₂	O ₁ L	LM	O ₁ O ₂	O ₁ M	ML
N	84	84	84	41	41	41
r	.83	.86	.90	.90	.69	.89
Av. Age at First Test (in years)	8.9	10.4	11.8	9.3	10.3	11.4
Av. Int. between Tests (in years)	1.6	1.3	1.2	1.0	1.1	1.2

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

The reliability coefficients for the original scale in both groups are above .80. So are the correlations between the two forms of the revised scales which have been considered here as analogous to reliability coefficients. The correlations between the original scale and each form of the revised, which give an estimate of the equivalence of ratings, are not significantly different¹ although the correlation with the M form is lower. Thus it can be said that the reliabilities for both scales and the correlations between both scales are essentially high and equal

Table II, which shows the correlations between the original scale and the two forms of the revised, and their respective reliabilities, using all the population that was available with no attempt to keep the composition of the groups the same, gives estimates that are all high with but one exception

TABLE II
Correlations Between the Various Forms of the Stanford-Binet Scales
for Populations of Variable Composition

Scales Correlated	O ₁ O ₂	L ₁ L ₂	M ₁ M ₂	OL	OM	LM	ML
N	118	85	44	146	61	116	89
r	.80 ¹	.85	.89	.82	.76	.88	.60 ¹
Av Age at First Test (in years)	10.3	9.7	10.0	11.8	12.1	10.9	10.3
Av Int between Tests (in years)	1.3	1.9	1.9	2.8	2.7	1.2	1.0

* An administration of an alternate form was included between the two forms correlated. Therefore these correlations are not between successive administrations as are the others

The estimate of correlation between the M form and the L form of the revised scale is significantly lower than any of the other estimates. However, since an estimate of this same correla-

¹ No P.E. is given in this study since an improved technique (Ridei, 10 pp 84-85) has been used to determine whether the difference between correlation coefficients is significant. The use of the P.E., is a crudely approximative method at best and in this particular case it is erroneous since the assumption of a normal distribution of correlation coefficients is probably violated in this case

ORIGINAL AND REVISED STANFORD BINET SCALES

tion obtained in Group B (Table I) is high, .89, the lower coefficient here, may be due to the particular sample taken.

Another view of the equivalence of ratings from scale to scale can be obtained from a study of the deviations from administration to administration. The results presented in Table II, where the grouping is according to I Q classification, show that within each scale and between scales the individuals with the lowest I Q's gain most upon retest, those of average I Q gain some, while those of highest I.Q actually lose.

TABLE III
Deviations Between Ratings in Successive Administrations According to I Q. Classifications

Scales Administered	O ₁ O ₂			LM or ML			OL or OM		
	Below 90	90 to 110	110 and above	Below 90	90 to 110	110 and above	Below 90	90 to 110	110 and above
I Q Level									
N	197	328	66	83	197	124	94	193	50
N (pos)	108	155	22	58	123	60	69	133	19
N (neg)	76	158	42	20	59	60	23	55	31
M (pos)	69	65	78	78	66	50	91	99	16
M (neg)	41	54	91	44	38	61	42	52	30
M	53	57	84	65	53	54	77	83	25
M	-22	+ 5	-32	14	130	- 5	56	46	- 13

We have here the expected tendency for the extremes of the distribution to migrate toward the mean with successive retestings (regression).

In Table IV where the deviations are classified according to the length of interval between tests, the mean of the absolute deviations increases as the interval becomes longer in every case but one

In that case, this reversal of tendency may be discounted in view of the small number of cases (10). It is concluded that the longer the interval between successive administrations, the greater the discrepancies in the ratings

Table V, which gives for both scales the relation between the size of deviation and the number of tests taken, shows that the mean of the absolute deviations decreases slightly with successive tests for the revised scale and does the same for the original scale with one exception

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE IV

Deviations in Ratings According to the Length of Interval Between Administrations

Scales Administered	O ₁ O ₂			LM or ML			OL or OM		
	Less than 1 year	1 year to 2 years	2 years and above	Less than 1 year	1 year to 2 years	2 years and above	Less than 1 year	1 year to 2 years	2 years and above
N	95	460	34	95	297	10	25	164	151
N (pos)	53	228	7	56	175	3	15	109	111
N (neg)	35	210	26	33	109	3	9	53	35
M (pos)	7.2	6.7	4.6	6.4	6.7	5.3	8.9	8.3	10.8
M (neg)	4.8	5.6	7.2	4.2	4.9	7.0	4.9	5.3	4.5
M	5.8	5.9	6.5	5.2	5.7	3.7	7.2	7.2	9.0
M	+2.2	+1.8	-4.6	+2.3	+4.6	-5	+3.6	+3.8	+6.9

TABLE V

Deviations in Ratings Between Successive Administrations in Relation to the Number of Tests Taken

Deviations	Original Scale				Revised Scale			
	O ₁ O ₂	O ₂ O ₃	O ₁ O ₃	O ₁ O ₄	I ₁ or M ₁	M ₁ L ₁	M ₂ or L ₁	L ₁ M ₂
N	244	166	103	64	133		69	
N (pos)	109	87	52	28	88		32	
N (neg)	124	65	48	34	38		33	
M (pos)	8.3	6.1	5.4	5.7	6.4		4.8	
M (neg)	6.1	5.8	4.7	5.2	5.4		3.9	
M	6.8	5.5	5.1	5.3	5.8		4.1	
M	+6.1	+9	+5	-1	+2.7		-1.3	

In general, with continued retesting, the discrepancies between successive administrations tend to become slightly smaller.

Table VI, which presents the deviations according to the chronological age of the individual at the time of the first test, shows a different trend of deviations with age for each scale

ORIGINAL AND REVISED STANFORD BINET SCALES

TABLE VI

Deviations in Ratings According to the Age of the Individual
at the First Test

Scales Administered	O ₁ O ₂			LM or ML			OL or OM		
	Below 8	8 to 10	Above 10	Below 8	8 to 10	Above 10	Below 8	8 to 10	Above 10
N	205	210	157	69	103	226	36	100	200
N (pos)	97	92	78	42	66	126	17	62	165
N (neg)	102	108	68	21	32	87	18	37	40
M (pos)	6.2	9.0	7.3	8.0	6.6	6.1	6.2	9.1	10.1
M (neg)	6.0	5.5	5.0	5.7	4.2	2.5	6.3	4.7	4.8
M	5.9	6.7	5.8	6.6	5.5	4.3	6.1	7.4	8.8
M	0	1.1	1.5	3.1	2.9	2.4	3	3.9	6.9

With the original scale, the mean of the absolute deviations is a maximum in the middle age group; with the revised scale, it decreases with age; and, between the original and revised scales, it increases with age. In considering the net gain upon retesting, it is found that for the original scale there is a small increase in net gain with age. For the revised scale there is a decrease in the amount of gain, and between the original and the revised there is a substantial increase in net gain with age. In no instance is there a net loss.

Changes in Dispersion

In studying changes in dispersion of the IQ distributions from test to retest in order to estimate how well the test will discriminate between members of a group upon retest, it was thought desirable to keep the population in any particular comparison constant to avoid any change in dispersion due to a change in the composition of the group. Four groups were used.

Groups A and C show the changes in dispersion on the original scale with one and two retests respectively; Groups B and D do the same for the revised scale. With successive administrations of the revised scales, the standard deviations decreased and in Group D this decrease was significant. This is what might be expected since there should be a regression toward the mean upon retesting. To the extent that there is regression a given test

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE VII

Changes in the Dispersions of I.Q. Distributions with Retesting

Group	A		B		C			D		
Scale	O ₁	O ₂	L ₁ or M ₁	M ₂ or L ₂	O ₁	O ₂	O ₃	L ₁ or M ₁	M ₂ or L ₂	L ₃ or M ₃
N	79	79	118	118	165	165	165	127	127	127
M	98.0	100.4	100.4	103.1	93.7	97.7	91.5	100.5	103.4	104.7
O	13.6	14.0	14.7	14.4	11.0	12.4	12.3	15.1	14.4	14.0

discriminates less well between the members of a group upon retest

However, in the original scale, the standard deviation on the first test distribution is smaller and significantly smaller than those of either the second or third administrations. This is true for both Groups B and D. These latter results may seem contrary to expectation, but it should be remembered that these retests occurred a year later on the average and thus the child was a year older. It is known that there is an increase in variability with mean test performance. In other words, as children grow older they tend to be more variable. The operation of this factor tended to mask the predilection for the distribution to regress toward the mean in the original scale, while in the revised scale the regression toward the mean was sufficiently great to obscure the opposite tendency. From a practical standpoint, then, it appears that with a given group, the discriminative ability of the original scale increases slightly upon retest while that of the revised scale decreases.

In the investigation of scatter, this term will be defined as the number of age levels through which an individual had to be tested to obtain his rating, from the level at which he passed all tests to and including the one at which he failed all. His scatter as defined above is larger than his range of successes by one age level. Scatter is used here as an approximate measure of the time taken to administer the test.² In general, the more levels over which an individual scatters, the longer it takes to administer the test.

The amount of scatter is limited by the number of age levels

² A more direct measure of the time required, such as the use of a stop-watch, could not be employed since this study was obtained from records which did not contain such information.

ORIGINAL AND REVISED STANFORD BINET SCALES

present in the test and by the age level at which the subject obtains his basal year score, i.e., an individual getting a basal score at Year XII on the original scale cannot scatter more than four test groups since there are no more. For this reason, the amount of scatter for both scales has been analyzed according to the basal year scores obtained. This arrangement makes explicit any limitation of scatter by the ceiling of the test.

TABLE VIII
Amount of Scatter Classified According to the Basal Year
Scores Obtained

Basal Year Score of Individuals		III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
Original Scale	N	68	46	97	119	157	130	91	129	-	41	-	0
	Mean	5.2	5.1	5.5	5.6	5.8	5.6	5.1	4.7	-	4.0	-	-
Revised Scales	N	12	19	25	77	68	115	87	81	79	58	62	106
	Mean	6.1	5.3	5.5	5.9	7.0	7.3	7.1	6.8	6.5	6.0	5.5	4.8

Table VIII shows that the scatter increases to a maximum in the middle range. The maximum scatter is at basal year VII for the original scale and basal year IX for the revised scale. It is at these points that the ceiling of the test begins to limit the amount of scatter. Since there are fewer test groups at the higher age levels in the original scale, it might be expected that this ceiling would make its influence felt earlier. This is the case. The revised scale has the greater scatter throughout, probably as a result of the increased number of tests in it. At basal age seven, this difference in scatter which has been only slight increases somewhat.

According to these results it would seem that the revised scale in general takes a longer time to administer. This is in agreement with the results reported by Krugman (6). No evaluation can be made of this finding, since it is not known to what extent the longer testing time results in increased accuracy of the rating obtained.

Inversions in Basal Year Scores

Inversions in basal year scores, i.e., instances in which an individual on a later test makes a lower basal year score than on his first, were studied since they cast some doubt upon the assumption that an individual would answer correctly all those items below his basal year level. In the original scales such inversions

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

occur in four per cent of the total possible instances (24 times out of 589) In the revised scale, they occur in nine per cent of the total possible instances (35 times out of 412) This difference is not statistically significant

Inversions in Item Level

Inversions in item level or skips, i.e., instances in which the individual failed all the tests at one age level but succeeded on one or more tests on a higher level were noted The assumption in testing is that the individual will not succeed in any test beyond the level at which he fails all This assumption and the pressing demand of time economy militate against testing the child beyond the level at which he fails all the tests so that these skips are not so frequent as they might otherwise be To the extent that the child is not given opportunity to perform on tests at a higher level where he sometimes achieves a random success, his rating is an underestimation of his true ability

In the original scale such skips occurred in four per cent of the tests (40 times out of a possible 958). In the revised scales they occurred a little less than one per cent of the time (eight times out of a possible 823) The difference between these proportions is significant. Apparently the grouping of tests, from the viewpoint of avoiding such skips, has been much better in the revised scales

Validity of Mental Year Groupings

To test whether the grouping of test items by mental years is such as to represent for each year's grouping the normal performance of children of that chronological age, the basal year scores were analyzed according to the average age of children achieving those scores Table IX, giving the mean chronological

TABLE IX
Average Ages of Children Making Various Basal Year Scores

Basal Year		III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
Original Scale	N	67	51	83	118	144	124	74	116		21		
	M	5.0	6.3	7.2	9.0	9.9	11.3	11.9	13.0		13.9		
	O	1.0	1.1	1.2	1.6	1.3	1.5	1.5	1.2		.8		
Revised Scales	N	9	15	23	72	57	107	87	81	71	58	64	92
	M	4.6	5.6	6.8	8.5	10.6	11.5	12.2	13.1	13.9	14.4	15.7	16.4
	O	.8	1.0	1.0	1.8	2.2	2.1	2.3	1.7	1.7	2.3	1.7	1.9

ORIGINAL AND REVISED STANFORD BINET SCALES

ages of children making the various basal year scores, shows that the mean ages are in every case significantly higher than the year level indicated by the score

Within the same basal year group, the mean age for the original scale is in no case significantly different from the mean age for the revised scale, indicating that though both scales do not meet the foregoing criterion for the grouping of test items, one is no better than the other

Summary

The data were gathered from 440 normal children who had taken a total of 958 original and 823 revised Stanford-Binet examinations. The results indicate that the reliabilities for both scales are high, over 80, and the correlations between scales are comparably high. In both scales, children with low I.Q. tend to gain more upon retests than do the children of average I.Q. while those of above-average I.Q. actually tend to lose upon retesting. For both tests, as the interval between successive administrations increases, so do the discrepancies between the test ratings. For both scales, as more tests are taken, the discrepancies between later tests tend to be smaller than those between earlier tests.

In the original scale the mean of the absolute deviations is a maximum in the middle age range, for the revised scale it decreases with age.

For the original scale there is a small increase in net gain with increasing age. In the revised scale there is a decrease in the amount of gain.

The dispersion of I.Q.'s and therefore the discriminative ability of the test increases with successive tests on the original scale; on the revised, however, the dispersion decreases.

The scatter on the revised scales is greater and reaches its maximum later than on the original scale.

Inversions in basal year scores are more frequent in the revised scale while skips are more frequent in the original scale. Basal year test groupings on either test do not represent the normal performance of children of the corresponding age but rather of children a year or two older.

BIBLIOGRAPHY

1. Berger, Arthur and Speevack, Morris. "An Analysis of the Range of Testing and Scattering Among Retarded Children on Form L of the Revised Stanford-Binet." *Journal of Educational Psychology*, XXXI, 1: 39-44.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

- 2 Bernicuter, Robert G and Carr, Edward J "The Interpretation of IQ's on the L M Stanford-Binet" *Journal of Educational Psychology*, XXIX, 312-14
- 3 Carlton, Theodore "Performance of Mental Defectives on the Revised Stanford-Binet, Form L" *Journal of Consulting Psychology*, IV, 2 61-5
- 4 Harriman, Philip Lawrence "Irregularity of Successes on the 1937 Stanford Revision" *Journal of Consulting Psychology*, III, 3 83 5
- 5 Hildreth, Gertrude "Retests with the New Stanford-Binet Scale" *Journal of Consulting Psychology*, III, 2 49-53
- 6 Krugman, Morris "Some Impressions of the Revised Stanford-Binet Scale" *Journal of Educational Psychology*, XXX, 8 594-603
- 7 Munson, Grace and Saffir, Milton A "A Comparative Study of Retest Ratings on the Original and Revised Stanford-Binet Intelligence Scales." Paper delivered at the American Psychological Association Meeting in California, 1940
- 8 Rheingold, Harriet L and Perce, Frances C "Comparison of Ratings on the Original and the Revised Stanford-Binet Intelligence Scales at the Borderline and Mental Defective Levels" *Proceedings from the American Association on Mental Deficiency*, XLIV (1939), 2: 110-19,
- 9 Rider, Paul R *An Introduction to Modern Statistical Methods* New York John Wiley & Sons, Inc., 1939.
- 10 Spearman, Charles "Measuring Intelligence—A Critical Notice" *Human Factor (London)*, XI (1937)
11. Terman, L. M. *The Measurement of Intelligence* Boston Houghton-Mifflin, 1916.
12. Terman, L. M and Merrill, M A *Measuring Intelligence* Boston Houghton-Mifflin, 1937

THE PREDICTION OF SCHOLASTIC SUCCESS IN A COLLEGE OF MEDICINE

DEWEY B STUIT

University of Iowa

The prediction of scholastic success in the professional colleges is a major personnel problem and one of primary significance to the individual, the colleges and society as a whole. Satisfactory achievement in the professional courses is the first step toward vocational success. Unless an individual can perform satisfactorily the work required for a professional degree, the question of ultimate vocational success need not be raised.

If the individual can be informed of his chances for success in a professional college before he enrolls it should be of great advantage to him in terms of time and money saved if he should otherwise fail. At the same time it should encourage those who possess the necessary ability to make the sacrifices which may be involved. The net results should be a better adjusted individual and a more competent profession. While these statements apply to all professional colleges, they seem particularly pertinent to medicine because the period of training is long, the expense to the individual is considerable, and the welfare of society demands highly competent medical men. The present investigation was undertaken to throw some light on the problem of predicting success in this professional area.

Specifically, it was the purpose of this study to investigate the value of liberal arts grade point averages and certain aptitude test scores as predictive indices of success in first year medicine at the State University of Iowa.¹ Because of the variations in grading standards at different institutions only those students who completed all of their undergraduate work at the University and who had complete records for one year of work in medicine were included in the study. Prior to 1938 standards for admission to the College of Medicine required at least two years or 60 semester hours of work in an approved college of arts and sciences; after 1938 this

¹The writer wishes to express his appreciation to Dean E. M. MacEwen of the State University of Iowa College of Medicine for making available the basic data and to Mr. C. William Applegate, research assistant in educational personnel, for his contribution to the statistical analyses made in the study.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

was changed to three years or 90 semester hours. Also in 1938 the required grade point average in liberal arts work was raised from 2.00² to 2.20. The restrictions imposed made it necessary to select students from entering classes as far back as 1934. The number and percentages selected from each class are presented in Table I.

TABLE I
Number and Percentage of Students Selected from Various
Freshman Classes in the College of Medicine

Year	Total Enrolled*	No. Used	Percentage
1934	113	34	30.08
1935	121	40	33.05
1936	112	21	18.75
1937	104	33	31.73
1938	55	14	25.45
Total	505	142	28.12

* The total number enrolled in each class includes students completing their liberal arts work at other institutions in whole or in part, those registered as freshmen for more than one year, and those who withdrew in the course of the year.

The predictive indices available for this group of 142 students included Iowa Qualifying Examination scores, Moss Medical Aptitude Test scores, and grade point averages for liberal arts work. The Iowa Qualifying Examination, administered to all entering freshmen, consists of the Iowa High School Content Examination, Iowa Silent Reading Test, Iowa Mathematics Aptitude Test, and the English Training Examination. A composite score, consisting of a weighted raw score total, is computed for the group of four examinations and is used as the score in the Iowa Qualifying Examination.

The purpose of this examination is to assist counselors in their advisory work with students and to predict the scholastic success of undergraduate students in various colleges and curricula. The Moss Medical Aptitude Test is administered to applicants for admission to colleges of medicine by the American Association of Medical Colleges. In considering the liberal arts work, the total grade point average, the "required science" and "total science" grade point averages were studied separately. Required science, as distinguished from total science, includes 32 hours of prescribed courses. The specific subjects prescribed in the liberal arts curricu-

² Grade point averages or point hour ratios are computed by considering A = 4, B = 3, C = 2, D = 1, Fd = 0.

PREDICTION OF SUCCESS IN A COLLEGE OF MEDICINE

lum are inorganic chemistry through qualitative analysis, quantitative analysis, elementary organic chemistry, elementary physics, and biological science, usually zoology.

The criterion of success in first-year medicine consisted of the student's grade point average at the close of the academic year. Makeups for subject conditions and incompletes were disregarded. It was felt that the grade first assigned in a course should be used because it represented a better appraisal of the student's performance in comparison with his fellows. The same practice was followed in computing grade point averages for the second year of work.

The raw scores in the aptitude tests had been converted to percentiles and were thus recorded. It was assumed that these percentiles were equivalent from year to year. For computational purposes the percentiles were converted into linear scores by the use of Hull's table.

Student Performance

The performance of the students in the aptitude tests and liberal arts work is shown in Table II. The mean linear score of 44.50 in the Moss Medical Aptitude Test is equivalent to a percentile score of about 40 on nation-wide norms. Data were also available for 240 additional students who did not meet all of the criteria used in the selection of the 142 students included in this study. It will be noted that the mean linear score for this group is slightly higher, but the range is almost identical. In the Iowa Qualifying Examination and its sub-tests, the group is definitely superior as indicated by the mean linear scores, but the range is very wide, varying from the seventh to the ninety-ninth percentile in the composite score. The mean grade point average of these students in liberal arts work is also definitely superior. The average point hour ratio for the College of Liberal Arts is about 2.20, while the students in this group achieved a 2.60 average. From these data one might conclude that the typical student who goes into medicine at Iowa is definitely superior in the Iowa Qualifying Examination and in his liberal arts work, but he may be somewhat below the average in the Moss Medical Aptitude Test.

The second phase of the study was concerned with the relationship between the various predictive indices and scholastic success in first year medicine. The coefficients of correlation expressing these relationships were computed by the product-moment method and are presented in Table III.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE II
Student Performance in the Aptitude Tests
and in
Liberal Arts Work

Predictive Index	N	M	σ	Range	
Moss Medical Aptitude	141	41.50	15.05	Linear Score	13.83
				Percentile	3.96
Moss Medical Aptitude	382*	46.00	16.05	Linear Score	0.87
				Percentile	0.98
Iowa Qualifying Exam.					
1 Composite Score	142	62.29	15.57	Linear Score	21.91
				Percentile	7.99
2 High School Content	142	62.90	15.85	Linear Score	9.91
				Percentile	2.99
3 Math Aptitude	142	62.65	15.94	Linear Score	21.91
				Percentile	7.99
4 English Training	142	56.59	15.65	Linear Score	9.91
				Percentile	2.99
5 Silent Reading	139	58.75	16.28	Linear Score	22.91
				Percentile	8.99
Required Science	142	2.62	4.53	1.50† 4.00	
Total Science	142	2.60	3.08	1.72† 4.00	
Total Liberal Arts	142	2.60	3.76	1.81† 3.61	
Fresh Med P.H.R.	142	2.36	619	0.53 3.84	
Fresh Med P.H.R.	382	2.33	643	0.53 4.00	
Fresh Med P.H.R.	112†	2.45	524	1.57 -3.84	
Soph Med P.H.R.	112†	2.16	528	1.03 3.63	

* A supplementary study was made of 382 students.

† Students of the group of 142 who completed two successive years

‡ The 2.20 requirement was in effect in 1938. Previous to 1930 this had been 1.50, and was then raised to 2.00. A few students admitted in 1930 did not enroll until 1934. Only one had a total grade point average below 2.00.

TABLE III
Correlation of the Predictive Indices with the Criterion

	N	r	P.E.
The Iowa Qualifying Examination			
Composite Score	142	.098	.056
High School Content Examination	142	.058	.056
Mathematics Aptitude Test	142	.108	.056
English Training Examination	142	.025	.056
Silent Reading Test	139	.075	.056
Moss Medical Aptitude Test	142	.226	.054
Moss Medical Aptitude Test	382	.316	.031
Liberal Arts Grade Point Averages			
Required Science	142	.419	.046
Total Science	142	.465	.045
Total Liberal Arts Work	142	.449	.045

PREDICTION OF SUCCESS IN A COLLEGE OF MEDICINE

Inspection of Table III reveals that the Iowa Qualifying Examination correlates very low with success in first year medicine, that the correlation between the Moss Medical Aptitude Test scores and the criterion is hardly significant and that the total science average is most closely associated with success in medicine as measured by first year grades. Examination of the scatter-diagrams provided several clues which may explain certain of the correlations. In the Iowa Qualifying Examination only 27 students received linear scores below 50, hence seriously restricting the range of talent of this group. As a result a majority of the students are concentrated in the first and second quadrants, those in the second quadrant having received high scores in the qualifying examination but achieving below average in first year medicine. Much the same picture is presented for each of the sub-tests comprising the qualifying examination. The data suggest a critical linear score of 40 or 45 in the composite score of the qualifying examination, for only eight students with qualifying scores below a linear score of 45 succeeded in making a 2.00 average or better in first year medicine.

The scatter-diagrams of the Moss Medical Aptitude Test present a striking contrast to those of the Iowa Qualifying Examination. A significant proportion of the students who score low in the test do very well in freshman medicine. As shown in Table II, the average grade in first year medicine is 2.36 and in the Moss test the mean linear score is 44.50. A total of 29 students or slightly over 20 per cent scored below average in the Moss test, but made grades above 2.40 in first year medicine. The student scoring lowest in the aptitude test succeeded in making a 2.60 grade point average. Poor performance in the Moss Medical Aptitude Test does not appear to indicate with a high degree of certainty that the student will do poorly in medicine at this institution. The scatter-diagrams for the 382 students present a similar picture.

Of the indices computed from the students' liberal arts records, the total science grade point average correlates best with scholastic success in medicine. However, there are some extreme deviates who reduce the magnitude of the coefficient of correlation. For example, one student with a 2.00 liberal arts record made a 3.35 grade point average in medicine while another with a 3.00 record in liberal arts work made only a 1.50 average in medicine. In general, however, there is rather close agreement between the grades in the two curricula. It does not appear that the liberal arts science record is superior in predictive capacity to the student's general average in undergraduate work. When the total science average and Moss

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Medical Aptitude Test scores are combined as predictive indices, the multiple correlation is .494. Apparently the Moss test does not add greatly to the predictive capacity of the science grades.

Supplementary evidence concerning the relation between the Moss Medical Aptitude Test scores and liberal arts grade point averages on the one hand and success in medicine on the other is furnished in Table IV. With one exception the mean grade point average in liberal arts work increases as the level of achievement in medicine increases. This is also true of the mean score in the Moss Medical Aptitude Test but the trend is not as pronounced. It is interesting to note that the range of performance as regards predictive indices is about the same for all levels of achievement. This makes the low correlations less surprising.

TABLE IV
Moss Medical Aptitude Test, Mean Linear Scores and Liberal Arts Mean Grade Point Averages for Various Levels of Achievement in Freshman Medicine

Freshman Medicine		Moss Aptitude			Required Science			Total Science			Total L. A. Work		
P.H.R.	N	M	σ	Range	M	σ	Range	M	σ	Range	M	σ	Range
3.00-4.00	24	52.17	14.93	22-83	2.93	51	2.00-4.00	2.92	49	2.00-4.00	2.93	35	2.16-3.61
2.50-2.99	34	40.59	14.61	13-73	2.66	39	1.88-3.38	2.61	34	2.00-3.22	2.60	21	2.11-3.13
2.00-2.49	42	47.60	13.06	19-71	2.65	40	1.63-3.50	2.63	28	2.00-3.22	2.65	22	2.05-3.50
1.50-1.99	30	40.28	13.22	15-68	2.37	35	1.50-3.00	2.32	33	1.72-3.07	2.35	14	1.81-3.31
0.00-1.49	12	37.17	15.92	19-76	2.25	18	1.75-2.75	2.29	16	2.00-2.67	2.28	16	2.00-2.58

TABLE V
Student Persistence in the College of Medicine at the State University of Iowa

Year Entered	One Year	Two Years	Three Years	Four Years
	N	N	N	N
1934	34	30	29	27
1935	40	35	32	32
1936	21	19	17	
1937	33	28		
1938	14			

In order to ascertain whether generalizations made concerning first year medicine would apply to other years, the freshman and sophomore grades for 112 students were correlated. The resulting coefficient was .722. It also seemed desirable to know if the students who complete one year of work continue beyond that point. The results are presented in Table V and seem to warrant the conclusion that the persistence of students beyond the freshman year is very high. It also appears that the first year's work is strongly indicative of later success in the medical school.

PREDICTION OF SUCCESS IN A COLLEGE OF MEDICINE

The results of the present study agree very well with those found at other institutions in this region. At Minnesota³ the correlation between Moss Aptitude Test scores and freshman honor points was found to be .27 for the class of 1938 and .22 for the class of 1939. Liberal arts grades for these same classes showed correlations with freshman grades in medicine of .57 and .46 respectively. In a study made of the classes entering the University of Illinois¹ in 1932 and 1933 the correlations between liberal arts averages and achievement in first year medicine were found to be .49 and .41 respectively. Comparable correlations for the liberal arts average in science were .57 and .42. The Moss Medical Aptitude Test was administered to the class entering in 1932 and correlated to the extent of .42 with first year medicine. Not all the reports on the prediction of success in medicine published in the *Journal of the Association of American Medical Colleges* are in agreement with these findings. Some report the Moss test as being superior to the liberal arts grade point average in predicting success in medicine while others find the reverse to be true. Apparently all agree, however, that aptitude tests and the undergraduate grade point averages furnish information which is valuable in selecting students for medical colleges.

Conclusions

The data seem to warrant the following conclusions for the population included in this study or populations which are similar.

1. Liberal arts grade point averages are the best predictive indices of success in first year medicine. Required science, total science and total liberal arts work are of about equal value in this respect.

2. The correlation between the Iowa Qualifying Examination scores and grades in freshman medicine is very low. However, the data do suggest a critical score which might be used by counselors in their advisory work with students who are interested in medicine as a career.

3. In this institution the Moss Medical Aptitude Test does not predict the student's level of achievement with high precision. Students scoring low in the test may do very well in medicine.

³J. W. Cavett, A. T. Henrici, and S. B. Lindley. "Tests of Medical Aptitude at Minnesota." *Journal of the Association of American Medical Colleges*, XII (September, 1937), 257-68.

¹George R. Moon. "Study of Premedical and Medical Scholastic Records of Students in the University of Illinois College of Medicine." *Journal of the Association of American Medical Colleges*, XIII (1938), 208-12.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

4 To predict success in medicine with greater accuracy will require tests which distinguish more clearly between various levels of ability. It is also possible that the criterion of success will need to be defined more precisely.

5. The counselor should not rely solely upon aptitude test scores and grade point averages in advising students about their probable success in medicine. Perhaps average achievement in aptitude tests and liberal arts work plus high interest and motivation will insure the student's scholastic success.

GIVEN AT THE UNIVERSITY OF CHICAGO[†]
A COMPARATIVE STUDY OF FRESHMAN WEEK TESTS

WILLIAM M SHANNER
Civil Aeronautics Authority
and
G FREDERIC KUDER
Social Security Board

One of the most crucial problems that confronts the educator of today is that of correctly advising students as to their education and vocational careers. In conjunction with this problem, educators and psychologists have constantly worked to secure more reliable and accurate information for use in counseling students. Almost every college and university now has an orientation week at which time all incoming students are required to take batteries of psychological and placement examinations, the results of which are used in advising students relative to their educational programs.

In September, 1938, a comprehensive battery of psychological and placement tests was administered the incoming freshman the relationship between these tests and succeeding academic class at the University of Chicago with a view toward studying achievement at the university. Among the tests administered the freshman group were the sixteen sub-tests of the American Council on Education Tests for Primary Mental Abilities, Experimental Edition, the 1938 Form, College Edition of the American Council on Education Psychological Examination, the College Entrance Examination Board's Scholastic Aptitude Examination, a physical sciences aptitude test, a social sciences aptitude test, Pressey's Special Reading Test, Form A; Pressey's Test on Reading Comprehension, Form A, and a vocabulary test. The physical sciences aptitude, social sciences aptitude, and vocabulary tests were locally constructed.

The first two years of the University of Chicago are devoted to a program of general education. The curriculum includes four introductory survey courses in the following fields: biological

[†] This study was made while the writers were with the Board of Examinations at the University of Chicago.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

sciences, humanities, physical sciences, and social sciences, and a number of elective second-year or sequence courses in specific subjects. The typical student takes two of the four general courses during his freshman year and the two remaining courses during his sophomore year. His program each year is completed by one or more of the other courses offered in the college. The general courses extend throughout the school year, and achievement is measured at the close of the year by means of a six-hour comprehensive examination. Attendance at a course is not required. The only requirement is the successful passing of the comprehensive examination. Many students are advised, upon the basis of their performances on the freshman week examinations, to attempt a comprehensive examination without taking the course. Since all students entering the University of Chicago as freshmen are required to take the four general courses, educational advisers are confronted with the problem of selecting the most appropriate general courses for the educational program of the student and advising him as to whether he needs additional assistance or should attempt the comprehensive examination without taking the course.*

By June, 1939, 501 of the freshmen entering the University in September, 1938, had taken one or more of the comprehensive examinations for the four survey courses and various sequence courses. The grades of the comprehensive examinations are reported in terms of derived scores having a mean of 20 and a standard deviation of 4. The average examination grade of each student was found by adding the derived scores for all his comprehensive examinations and dividing by the number of examinations.

Test Scores and Average Grades

Table I reports the correlation between the various freshman week tests and average examination grades. The testing time of each examination is also reported. The social sciences aptitude test has the highest correlation with average grades (.575). The physical sciences aptitude, the American Council Psychological Examination, and the College Entrance Examination Board's

* For a comprehensive description of the organization of the first two years of the University of Chicago, see Chauncey Samuel Boucher, and A. J. Brumbaugh, *The Chicago College Plan* (Chicago: The University of Chicago Press, June, 1940), pp. xii-413.

A STUDY OF FRESHMAN WEEK TESTS

Scholastic Aptitude Examination have just slightly lower correlations. All these tests, with the exception of the CEEB Scholastic Aptitude, require approximately one hour of testing time each, the CEEB examination requires two hours' time.

The social sciences aptitude test is essentially a reading test. It consists of three selections, one each drawn from the fields of economics, sociology, and political science. Each paragraph is followed by a number of questions based upon an understanding of the materials covered. The test is a revision of a test given experimentally the previous year.

The physical sciences aptitude test consists of (1) a section on vocabulary in the field, (2) questions involving the interpretation of mathematical formulas, and (3) a reading test containing chemistry and physics selections. It is the product of a process of analysis and revision carried out over a period of years.

The results of the 16 tests of the Primary Mental Abilities battery are reported in terms of seven composite scores, each an approximation to a factor. Scores for the following abilities are reported for the test:

Perceptual. This ability, measured by the verbal enumeration and identical forms tests, may be described as one's facility in finding detail which is significant to him or detail which he is seeking.

Number. This factor consists of facility with simple numerical work and is measured by the tests of rapid addition and multiplication.

Verbal. The verbal factor manifests itself in the completion and in same or opposite tests. It is roughly the ability to deal readily and quickly with verbal materials.

Spatial. The spatial factor is measured by tests requiring the subject to think visually of geometric forms and of objects in space.

Memory. This factor is one's ability to memorize various materials. One test requiring the memorization of initials with names, and a second test requiring the association of words with numbers are used in measuring the ability.

Inductive Reasoning. The induction factor may be described as one's ability to discover some rule or principle in various arrangements of material. A numerical, a verbal, and a spatial test are used in estimating the ability.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Deductive Reasoning. The deductive factor may be described as facility in formal reasoning. It is measured by tests of arithmetic problems, number series, and perception of mechanical movements.

The largest correlation between a composite primary ability score and average grades is for the verbal composite (.415) which requires 16 minutes of testing time. The remaining correlations are much lower. These results are reasonable, since the composite scores represent specific abilities, and average grades represent general academic proficiency.

Test Scores and Course Grades

Table II reports the correlation between various freshman week tests and grades for the four general courses. The two largest correlations (.654 and .648) are for the physical and social sciences aptitude tests with the respective general courses. The correlations for the American Council Psychological and the CEEB Scholastic Aptitude Examination show no statistically significant difference for the biological sciences, social sciences, and physical sciences general course examinations. However, the correlation between the CEEB and humanities is significantly greater than between humanities and the American Council Psychological Examination. It is of interest to note the variations in the size of the correlation coefficients for the composite scores of the Primary Mental Abilities battery. The two largest correlations for the composite scores are between Deduction and grades in the physical sciences, and between Verbal and grades in the humanities. One might very well expect these phenomena. At the same time, humanities shows a correlation of only .071 with the composite Spatial score.

The degree of independence of the seven composite scores of the Primary Mental Abilities battery is reported in Table III, which gives the intercorrelations for the seven scores. Slightly over half of the correlations in the table are less than .300 and two can be considered as zero. These small correlations suggest a considerable degree of independence for these scores and that they might well measure specific abilities. The intercorrelations among Spatial, Induction, and Deduction scores are all very near .500 and thus give evidence of considerable dependence of scores.

The first line of Table IV reports the multiple correlation coefficients for the combination of the two primary ability composites having the highest validities with respect to each of the four general

A STUDY OF FRESHMAN WEEK TESTS

courses The Verbal and Deduction scores, from tests requiring 70 minutes, were combined for all except the humanities course. For this course the Verbal and Number scores, from tests requiring 35 minutes, were combined. The second line of Table IV reports the multiple correlations obtained by combining all seven scores of the Primary Mental Abilities Tests These coefficients are not markedly higher than those obtained from the best two in each case

TABLE I

Testing Time Required for Administering Various Psychological and Placement Tests to the 1938 Freshman Class at the University of Chicago and the Correlation of these Tests with Average Grades

Test	Testing Time in Minutes	Correlation with Average Grades	Test	Testing Time in Minutes	Correlation with Average Grades
Perception*	20	117	American Council Psychological Examination	56	523
Number*	19	310	College Entrance Examination Board's Scholastic Aptitude	120	542
Verbal*	16	415	Physical Sciences Aptitude	60	522
Spatial*	33	184	Social Sciences Aptitude	60	575
Memory*	33	204	Pressey's Special Reading	60	477
Induction*	46	229	Pressey's Reading Comprehension	†	326
Deduction*	54	378	Vocabulary	25	486

* Composite scores for the Thurstone Tests for Primary Mental Abilities

† Specific time limits are not given; the students are given the time necessary to read entire reading selection and answer the questions. Approximately 20 minutes are required.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE II

Coefficients of Correlation between Various Psychological and Placement Tests and the Four Introductory General Courses at the University of Chicago

Test	Biological Sciences	Humanities	Physical Sciences	Social Sciences
Perception*	.084	.129	.166	.135
Number*	.207	.265	.272	.300
Verbal*	.380	.472	.476	.435
Spatial*	.225	.071	.139	.131
Memory*	.145	.127	.177	.160
Induction*	.216	.029	.247	.196
Deduction*	.418	.190	.485	.427
American Council Psychological Examination	.483	.485	.482	.569
College Entrance Examination Board's Scholastic Aptitude Examination	.479	.544	.471	.577
Physical Sciences Aptitude	—	—	.654	—
Social Sciences Aptitude	—	—	—	.648

* Composite scores for the Thurstone Tests for Primary Mental Abilities.

TABLE III

Intercorrelations for the Seven Composite Scores for the Thurstone Tests for Primary Mental Abilities

	Number	Verbal	Spatial	Memory	Induction	Deduction
Perception	.237	.371	.392	.046	.355	.153
	Number	.250	.204	.188	.306	.336
		Verbal	.218	.156	.347	.368
			Spatial	.077	.490	.475
				Memory	.170	.126
					Induction	.533

A STUDY OF FRESHMAN WEEK TESTS

TABLE IV

Multiple Correlation Coefficients between Various Combinations of the Composite Scores of the Primary Mental Abilities Tests and the Four Introductory General Courses

Combination of Composite Scores	Biological Sciences	Humanities	Physical Sciences	Social Sciences
Two Best Predicting Composite Scores	.484	.496	.529	.521
All Seven Composite Scores	.500	.541	.561	.556

Conclusion

Two rather striking observations may be made on the basis of the results reported

- (1) Marks in the four courses can be predicted by combining two fairly short primary abilities measures about as well as by using the one-hour American Council Psychological Examination or the two-hour scholastic aptitude test of the College Entrance Examination Board, both of which were constructed for the purpose of predicting scholarship. This result is the more remarkable since the Primary Abilities Tests were not specifically constructed for the purpose of predicting grades.
- (2) Tests developed for the specific situation are in the present case more efficient prognostic measures than any other single test or combination of tests studied. The validities of the aptitude tests for the physical sciences and the social sciences are significantly higher than the other validities obtained.

These two results appear to be essentially contradictory. One of them argues for the development of a number of relatively independent measures and the use of those which, in combination, are most efficient for the prediction of any selected criterion or group of criteria. The other seems to indicate that tests constructed and revised in the light of analysis with respect to the local situation are most effective, at least when compared with two of the better scholastic aptitude tests constructed for general use. This conclusion is valid for the tests studied in their present state of development. However, it is apparent that what can be measured in composite tests, such as the aptitude tests in the fields of the social and

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

physical sciences, can be measured by a number of more specific and relatively independent measures. The difference between the predictive efficiency of the primary abilities measures used as compared with the specific aptitude tests must be attributed to the fact that the former do not sample some of the attributes included in the latter. The fact that the Primary Mental Abilities Tests in combination produce fairly high validities although they were not developed for the purpose of predicting scholarship is indicative of the promise in this type of measurement. As the experimental tests of primary abilities are perfected and expanded to include other abilities involved in scholastic success, it is reasonable to expect combinations of them to approach and equal the validities of tests constructed for each of a number of specific situations. This development will make practical a much more efficient use of test material when a number of criteria are to be predicted.

NOTE ON A SIMPLIFIED METHOD OF COMPUTING TEST RELIABILITY

C J HOYT

University of Minnesota

Kuder and Richardson¹ have presented the theoretical background as well as useful formulas for a new and improved procedure for estimating the coefficient of test reliability. In a later paper² they have labeled their procedure "the method of rational equivalence." Their results appear to have a number of important advantages over the split-half correlation method used in conjunction with the Spearman-Brown formula. With the split-half method the obtained coefficient may be an overestimate or an underestimate of the actual reliability. With the method of rational equivalence the estimate derived is known to be never an overestimate.³ This fact alone is sufficient for recommending the displacement of the split-half procedure, although there are other advantages, as pointed out below.

The theoretical soundness of the Kuder-Richardson derivation is indicated by the fact that analysis of variance techniques applied to this problem produce an identical formula. The present writer's derivation, using an approach entirely different from that used by Kuder and Richardson, will appear elsewhere.

The use of the formula recommended by the authors for general use requires only the same primary data as are ordinarily obtained in a careful analysis of a test. Consequently, it is not necessary to obtain the scores on separate parts of the test. The possibility of obtaining varying results with different methods of dividing the test is also obviated. The computations involved

¹G. F. Kuder and M. W. Richardson "The Theory of the Estimation of Test Reliability" *Psychometrika*, II (1937), 151-60.

²M. W. Richardson and G. F. Kuder "The Calculation of Test Reliability Coefficients Based on the Method of Rational Equivalence" *Journal of Educational Psychology*, XL (1939), 681-87.

³This statement is strictly true for the population used. Sampling errors are, of course, not eliminated. For a discussion of sampling errors the reader is referred to Robert W. Jackson, "Reliability of Mental Tests" *British Journal of Psychology*, XXIX (1939), 267-87.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

in computing the coefficient of reliability by the method of rational equivalence can be performed in a few simple steps that do not require any special statistical knowledge

The increasing use of the method of rational equivalence for the estimation of test reliability leads the writer to describe a procedure which he has found to be particularly efficient. Although Kuder and Richardson present a number of formulas involving various degrees of rigor, they recommend their formula (20) for general use. Their empirical findings and those of a number of others who have been using the method indicate that the results obtained from their formula (20) closely approximate those obtained by the more rigorous formulas. The steps outlined below have therefore been developed for a variant of the recommended formula.¹

1. Score the tests for the number of right answers. Obtain the sum of these scores for all the subjects. This value is T in formula (1) below.
2. Square each of these scores and obtain the sum of these squares for all the subjects. This sum is Ss in the formula below.
3. Make a tally of the test responses to each item and obtain the count of the number correct for each item. The total of these counts should equal the T obtained in step 1.
4. Square the count obtained for each item and obtain the sum of these squares for all the items. This sum is Si in the formula below.
5. Using the values obtained in the steps above, solve the following formula for r_{tt} , the reliability of the test. In this formula, k is the number of subjects taking the test and n is the number of items in the test.

$$r_{tt} = \frac{n}{n-1} \frac{kSs + Si - T(T + k)}{kSs - T^2} \quad (1)$$

In the use and analysis of a test some of these steps will already have been performed. Use of the item counter of the International Scoring Machine will greatly facilitate step 3. If

¹ While the present paper was in press, Paul L. Dressel published other variants of the Kuder-Richardson formulas. Formula (1) above is equivalent to formula (4) of Dressel's Paper. Paul L. Dressel, "Some Remarks on the Kuder-Richardson Reliability Coefficient" *Psychometrika*, V (1940), 305-10.

A METHOD OF COMPUTING TEST RELIABILITY

a computing machine is not available, steps (2) and (4) can be greatly facilitated by the use of a table of squares and an ordinary adding machine

The use of formula (1) will be illustrated in a particular example involving a test of 250 items administered to a group of 33 students in the College of Pharmacy at the University of Minnesota. The values obtained in this case were as follows

$$\begin{aligned} S_1 &= 112,873 & T &= 4829 & S_2 &= 727,351 \\ r_{11} &= \frac{250}{249} \frac{33(727,351) - 112,873 - 4829(4829 - 1) - 33}{33(727,351) - (4829)^2} \\ &= \frac{250}{249} \frac{636,858}{683,342} = \frac{159,214,500}{170,152,158} = .936 \end{aligned}$$

Formula (1) is algebraically equivalent to formula (20) presented by Kuder and Richardson. Their formula (20) is as follows:

$$\begin{aligned} r_{11} &= \frac{n}{n-1} \frac{\sigma_t^2 - npq}{\sigma_t^2} \\ &= \frac{n}{n-1} \frac{\sigma_t^2 - \sum p_i q_i}{\sigma_t^2}, \end{aligned}$$

where σ_t is the standard deviation of the distribution of test scores, p_i is the proportion of students passing each item taken in turn, and q_i is the proportion failing that item.

It should be remembered that this procedure is no more applicable to speed tests than is the Spearman-Brown formula.

MEASUREMENT ABSTRACTS

Bedell, Ralph "Scoring Weighted Multiple Keyed Tests on the IBM Counting Sorter" *Psychometrika*, V (1940), 195-201

Tests or personal inventories with differential item response weights may be scored by means of punch card equipment. Detailed instructions are given for preparing the cards and scoring the forms. The scoring speed is approximately four to eight times that attained by manual scoring. (Courtesy *Psychometrika*.)

Blakey, Robert, "A Re-Analysis of a Test of the Theory of Two Factors" *Psychometrika*, V (1940), 121-36

The study of William Brown and William Stephenson, "A Test of the Theory of Two Factors," is re-analyzed by means of the Thurstone multiple factor methods. No tests or correlations are left out of the original table of correlations as is done in the original analysis in an attempt to validate the two-factor theory. Space, verbal, and perceptual speed factors similar to those found by Thurstone, Wright, and Garrett are identified. A common factor of "Maturation" is postulated to account for the remaining communality of the tests. A fifth factor is considered to have no significance due to the small amount of variance which it contributes to the total. (Courtesy *Psychometrika*.)

Blum, M. L. "A Contribution to Manual Aptitude Measurement in Industry: the Value of Certain Dexterity Measures for the Selection of Workers in a Watch Factory" *Journal of Applied Psychology*, XXIV (1940), 381-416

Job analysis of watch assembling suggested the importance of the ability to make fine finger movements, the ability to handle tweezers, and the ability to continue to perform delicate tasks without increasing tension or maladjustment. Three criteria of proficiency were established: length of employment, salary ratio, and foremen's ratings. Two hundred and fifty-eight women (37 workers, 137 applicants before being hired, 84 applicants after being hired) were examined with the O'Connor Finger Dexterity and Tweezer Dexterity tests. Time scores showed the highest prediction of the proficiency criteria. The practical value of critical time scores on the dexterity tests was indicated. W. A. Varvel.

MEASUREMENT ABSTRACTS

Cattell, R B "A Culture-Free Intelligence Test " Part I *Journal of Educational Psychology*, XXXI (1940), 161-79

A common source of error in the Binet-Simon type of test arises from the influence of academic experience and general cultural background. Instead of sampling the "common knowledge" of the subject, the test emphasizes the perception of relations inherent in objects and processes common to a wide range of cultural groups. One hundred multiple choice scaled items are chosen from mazes, series, classifications, progressive matrices (3 types), and mirror images. The progressive matrix test consists of combined analogy and progressive series items. *Harold Bechtoldt*.

Dunlap, Jack W "Problems Arising from the Use of a Separate Answer Sheet," *Journal of Psychology*, X (1940), 3-48.

The use of a separate answer sheet has been considered in terms of validity and reliability of the more conventional type of response. Underlining, marking parentheses, marking separate answer sheets using serial and repetitive numbering for choices with the answer sheets of both articulated and non-articulated types lead to practically identical results. Comparisons were made in terms of means, standard deviations, reliabilities, and validity of test results for both fourth- and eighth-grade pupils. The use of an articulated, serial numbered answer sheet is recommended for tests short enough for all answers to be recorded on a single side of the sheet. *Harold Bechtoldt*

Harrell, Willard "A Factor Analysis of Mechanical Ability Tests" *Psychometrika*, V (1940), 17-33.

The intercorrelations of 37 variables, including the Minnesota battery of "mechanical ability" tests, the seven MacQuarrie tests of "mechanical ability," O'Connor's Wiggly blocks, and the Stenquist picture-matching test, were analyzed by Thurstone's centroid method. Five factors, Perceptual, Verbal, Youth, Manual Agility, and Spatial, were taken out. Factors prominent in so-called mechanical ability tests are the Spatial and Perceptual ones with MacQuarrie's dotting test significantly high in the Manual Agility factor. Each of the factors can be measured with group pencil-and-paper tests. (Courtesy *Psychometrika*.)

Harrell, T. W. and Faubion, R W "Selection Tests for Aviation Mechanics," *Journal of Consulting Psychology*, IV (1940), 104-05.

Students of the United States Army Air Corps Technical Schools take a basic course of Shop Mathematics, Mechanical

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Drafting and Blueprint Reading, Air Corps Fundamentals, Metal Work, and Electricity besides specializing in some field. Correlations of 38 tests with these five basic courses range from -20 to $+54$. Four tests give a multiple correlation of $.72$ with a composite basic grade. A factor analysis is being made of 24 of these variables. *Harold Bechtoldt*

Johnson, A. P. "A Study of One Company's Criteria for Selecting College Graduates." *Journal of Applied Psychology*, XXIV (1940), 253-64.

A company had for some years considered applications for sales positions (personnel, advertising, and sales promotion) on the basis of an intelligence test, a vocabulary test, and ratings on family background, industriousness, extroversion-introversion, and flair for writing. The present study examines the data for 80 applicants (41 hired, 39 rejected) and seeks to objectify the ratings and to establish estimates of their reliability and validity. The combined ratings of six members of a class in industrial psychology showed satisfactory reliability. Ratings on "writing flair" most markedly differentiated the hired from the rejected. Ratings on "family background" showed the highest correlation ($+0.40 \pm 0.12$) with service or merit ratings made by five company executives on 23 workers. *W. A. Varvel*

McCloy, C. H. "The Measurement of Speed in Motor Performance." *Psychometrika*, V (1940), 173-82.

When the centroid method of factor analysis was applied to two sets of data on athletic performances, three significant factors emerged: strength, velocity, and dead weight. Scores on this speed factor were predicted by the multiple regression technique, the factor loadings on the speed factor being used as the criterion correlations, and these predicted scores were correlated with each of the other variables. When the original tables, augmented by the new speed variable, were refactored, the computed speed factor fell on the speed axis as a primary trait. It is thus shown that it is possible to isolate and measure a factor which appears in variables under consideration only as a compound. (Courtesy *Psychometrika*)

Palmer, C. E., and Klein, H. "A Table of the Double Integral of the Gaussian Probability Function." *Child Development*, XI (1940), 61-8. *F. A. Kingsbury*

MEASUREMENT ABSTRACTS

Roslow, S., Wulfeck, W. H., and Corby, P. G. "Consumer and Opinion Research: Experimental Studies on the Form of the Question" *Journal of Applied Psychology*, XXIV (1940), 334-46

Summaries of the results of eight studies on varying the form of questions are given. Alternate forms of the questionnaire, successive forms one month apart, and free response questions were among the methods used. The use of stereotypes or emotionally charged words produced significant changes in responses. Slight changes in wording may or may not result in changes in frequencies of the response choices. The completeness and number of alternatives offered in check lists tend to influence the proportions for any one response, while the results from free-response questions may be definitely misleading. *Harold Bechtoldt*

Sarbin, T. R., and Berdie, R. F. "Relation of Measured Interests to the Allport-Vernon Study of Values" *Journal of Applied Psychology*, XXIV (1940), 287-96

Fifty-two university students were given the Allport-Vernon Scale and the Strong Vocational Interest Blank, Form M. A modification of the pattern analysis described by Darley was applied to the Strong profiles. Occupational keys were grouped according to the results of factor analysis studies. "A few of the occupational groups showing measured interest patterns are characterized by certain profiles on the Allport-Vernon Scale." Although there is considerable overlapping between groups, "it is possible, nevertheless, . . . to use the Allport-Vernon Scale to approximate certain occupational interest types as measured by Strong. Thus, a definite but limited use is demonstrated for the Allport-Vernon scores when it is desirable to distinguish or identify vocational interest types in the professional, sales, or 'uplift' occupations." *W. A. Varvel*

Schultz, R. S. "Preliminary Study of an Industrial Revision of the Revised Minnesota Paper Form Board Test" *Journal of Applied Psychology*, XXIV (1940), 463-67

The Likert-Quasha Revised Minnesota Paper Form Board Test (Form AA) was further revised for industrial use in order to decrease the demand on verbal comprehension of the instructions and to simplify the response required. A preliminary study of this industrial revision is reported. Correlations with the Revised Minnesota range from +.71 to -.86. Scores on the industrial revision tend to be significantly higher. Twenty-one engineering students obtained a higher average score than did 42

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

trade-school boys and 57 high-school girls. Correlations with intelligence correspond with those found in previous studies with the Paper Form Board Test. *W. A. Varvel*

Seder, M. "The Vocational Interests of Professional Women" Part II. *Journal of Applied Psychology*, XXIV (1940), 265-72

Sixty women physicians and 69 life insurance saleswomen filled out both the men's form and the women's form of the Strong Vocational Interest Blank. For the 268 items common to the two forms, the median number of discrepant responses was 18 per cent, so that the test-retest reliability is considered satisfactory. The common items are as heavily or more heavily weighted than items occurring on only one blank. In general there is substantial agreement between the weights assigned to the response to each item by the men's key and by the women's key for the same occupation. "All indications of this study are that differences between sexes in an occupation are usually less frequent and less important than similarities." It is suggested that a common blank should be composed and that where sex differences actually appear an occupational key for each sex should be constructed. *W. A. Varvel*

Thurstone, L. L. "Experimental Study of Simple Structure." *Psychometrika*, V (1940), 153-68

A battery of 36 tests was given to a group of high-school seniors. The factorial analysis reveals essentially the same primary factors that were found in previous studies. The test battery reveals a simple structure. (Courtesy *Psychometrika*)

Tucker, Ledyard R. "The Role of Correlated Factors in Factor Analysis." *Psychometrika*, V (1940), 141-52.

The fundamental factor theorem is developed in matrix form for the case of correlated factors. The properties of the correlated factor system are discussed, and some effects of sampling error considered. The psychological meaning of correlated factors is discussed, and several mechanisms by which general factors may operate in the factorial system are indicated. (Courtesy *Psychometrika*.)

Walker, Helen M. "Degrees of Freedom." *Journal of Educational Psychology*, XXXI (1940), 253-69

The number of degrees of freedom is a basic concept in small sample theory. Most textbooks omit a discussion of this topic, and many texts give incorrect formulae and procedures because of ignoring it. The development starts with the freedom of move-

MEASUREMENT NEWS

ment of a point in space under certain restraining conditions and utilizes the representation of a statistical sample by a single point in N-dimensional space. Illustrations are presented showing how to determine the number of degrees of freedom appropriate for use in certain common situations, as standard error of the mean, Chi-square test, contingency tables, partial correlation, and analysis of variance formulae. *Harold Bechtoldt.*

Young, P. V. "The Validity of Schedules and Questionnaires," *Journal of Educational Sociology*, XIV (1940), 22-6

A brief summary is given of an experiment with a variety of questionnaires and schedules as used on three considerably homogeneous communities. Objective, quantitative data were difficult to obtain. The data shed little light on complexities of social patterns or on behavior patterns of cultural worlds in relation to social life and personality adjustment. A review is given of some problems involved in the construction of such instruments and of circumstances when they can most advantageously be used. *Calvin Taylor*

| 4 |

MEASUREMENT NEWS*

A Personnel Research Section has recently been established in the War Department under the Adjutant General. The function of the section is to devise and assemble procedures for the classification of military personnel. Dr. W. V. Bingham is the director of this section. Among the professional members of the staff are Dr. T. W. Harrell, on leave from the University of Illinois, Mr. W. M. Shanner, on leave from the University of Chicago, and Dr. Willis Schaefel, formerly of the University of Chicago. This section has the technical advice of a National Research Council committee on the Classification of Military Personnel. Members of the committee are

Drs. Walter V. Bingham, Carl C. Brigham, Princeton University, Henry E. Garrett, Columbia University, L. J. O'Rourke, United States Civil Service Commission, Marion W. Richardson, United States Civil Service Commission, Carroll L. Shartle, Social Security Board, and L. L. Thurstone, University of Chicago.

As a part of the national defense program the Occupational Analysis Section of the United States Bureau of Employment Security is making job analyses of occupations in the United

* Notes for this department should be sent to Dr. M. W. Richardson, United States Civil Service Commission, Washington, D. C.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

States Army Over seven thousand analyses will be made and job specifications will be prepared to aid the Army in making its assignments of personnel The Army is also using the Oral Trade Questions which have been developed by the Employment Service, as well as the recently published "Dictionary of Occupational Titles" New aptitude tests developed by the Occupational Analysis section are being released to both the Army and the Navy

Another field of activities of the Occupational Analysis Section is that of assisting local employment offices to select rapid learners for defense jobs New aptitude tests are being developed for this task New trade tests are also being constructed for defense jobs requiring highly skilled workers The greatest attention is being given to those jobs which are important both to the armed forces and to the civilian defense industries

The Occupational Analysis Section is under the supervision of Dr C L Shartle

The Washington Psychometric Society was organized November 13, 1940, with a charter membership of eleven The following officers were elected: M W. Richardson, president, N J Van Steenberg, secretary, C. R. Brolyer, treasurer It is planned to hold meetings once a month

Machine methods as applied to the field of measurement formed the major subject of discussion at an "Educational Research Forum" held at the Homestead of the International Business Machines Corporation at Endicott, New York, during the week of August 26 to 31 A limited number of transcripts of the proceedings are available to those interested Requests for transcripts should be sent to Mr. E C Schroedel, Manager, Institutional Department, International Business Machines Corporation, 590 Madison Avenue, New York, New York

The papers presented at the Forum are listed below

Computation of Statistical Constants

"The Value of the Collator in Using Prepunched Cards for Obtaining Moments and Product Moments."—Alan D Meacham

"The Computation of Means, Standard Deviations and Correlations by Use of the Tabulator When the Numbers are Either Positive or Negative"—Jack W Dunlap

"Summary of Problems in Computation of Statistical Constants"—Paul S Dwyer

MEASUREMENT NEWS

"The Design of Tabulating Procedures in Relation to Automatic Error Control in Statistical Analysis"—Charles R Langmuir

"Code Numbers and Coding as Aids to Research"—Herbert A Toops

Classification and Prediction

"Four Aspects of Factor Analysis A Problem for Which Machine Procedures are Needed"—Harry H Harman

"Use of Tabulating and Scoring Machines in Factor Analysis"—Ledyard R Tucker

"Canonicals"—Irving Lorge

"A Successive Approximation Solution for Prediction Problems Involving a Large Number of Variables"—John C Flanagan

"Problems of Classification of Personnel in the Army"—Truman L Kelley

"Army Testing Problems."—T W Harrell

Test Construction

"Computing Difficulty Index and Validity Index in Item Analysis by IBM Machines"—John M Stalnaker

"Item Analysis by Test Scoring Machine Graphic Item Counter."—John C Flanagan

"Repetitive Scoring of Interest and Personality Tests in Developing Item Weights by an Iterative Process"—Robert T. Rock, Jr

Testing Programs

"The Integration of the Test Scoring Machine with Tabulating Equipment in a System of Progress Tests and Comprehensive Examinations."—J V McQuitty

"Applications of Electric Accounting Machines in Reporting Individual and Group Results in a Testing Program"—Charles R Langmuir

"The Facilitation of the Analysis and Distribution of College Entrance Test Data in a Statewide Testing Program."—E. L. Stromberg

In the spring of 1939, at the request of a number of school teachers and administrators throughout the United States, the American Council on Education appointed the National Commit-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

tee on Teacher Examinations, and authorized it to supervise and delegate to the Cooperative Test Service of the American Council the task of preparing a battery of objective tests for the examination of teaching candidates. The National Teacher Examinations were administered for the first time in various centers throughout the United States on March 29-30, 1940.

New editions of the Teacher Examinations are being prepared for administration in 1941. The tests cover such areas as understanding and use of the English language; reasoning ability; knowledge of contemporary affairs, general cultural information, understanding of professional educational points of view, goals, attitudes, and methods, and mastery of subject matter to be taught. All examinations are objective, consisting of short answer items involving multiple choice response. In 1941 the National Teacher Examinations will be administered two full days. Approximately twelve hours of testing time are required for the examinations.

The dates which have been named by the National Committee for the administration of the Teacher Examinations in 1941 are March 14 and 15.

The examinations have, of necessity, been limited to intellectual, academic, and cultural materials. Other important factors that determine teaching success, such as training, experience, personality characteristics, social adaptability, and others are judged independently by the local authority to whom the candidate applies.

A revised series of Cooperative General Achievement Tests were introduced this fall by the Cooperative Test Service. The revised series beginning with Form QR includes Test I: A Test of General Proficiency in the Field of Social Studies, Test II: A Test of General Proficiency in the Field of Natural Sciences, and Test III: A Test of General Proficiency in the Field of Mathematics. These general proficiency tests are not composed of questions dealing with topical content of the fields covered. Instead, each test is divided into two parts: the first, testing for knowledge of the terms and concepts essential to an understanding of the area in question, the second, testing the student's ability to comprehend and interpret typical materials in the fields.

PRIMARY MENTAL ABILITIES OF CHILDREN¹

HILMA G. THURSTON

Chicago Teachers College

FOR MANY years psychologists have been accustomed to the problems of special abilities and disabilities. These are, in fact, the principal concern of the school psychologists who deal with children who cannot read, have a blind spot for numbers, or do one thing remarkably well and other things poorly. It seems strange with all this experience in differential psychology that we have clung so long to the practice of summarizing a child's mental endowment by a single index, such as the mental age, the intelligence quotient, the percentile rank in general intelligence, and other single average measures. An average index of mental endowment should be useful for many educational purposes, but it should not be regarded as more than the average of several tests. Two children with the same mental age can be entirely different persons, as is well known. There is nothing wrong about using a mental age or an intelligence quotient if it is understood as an average of several tests. The error that is frequently made is interpreting it as measuring some basic functional unity when it is known to be nothing more than a composite of many functional unities.

The researches on the primary mental abilities which have been in progress for several years have had as their first purposes the identification and definition of the independent factors of mind. As the nature of the abilities became more

¹The studies reported in this paper have been carried out under the joint sponsorship of the Chicago Public Schools, the University of Chicago, and the American Council on Education.

clearly indicated by successive studies, a second purpose of a more practical nature has been involved in some of the studies. This purpose has been to prepare a set of tests of psychological significance and practicable adaptability to the school testing and guidance program. The series of studies will be summarized in this paper, the battery of tests soon to be available will be described, and some of the problems now being investigated will be discussed briefly.

Previous Studies

The first study in this series involved the use of 56 psychological examinations that were given to a group of about 250 college students. That study revealed a number of primary abilities, some of which were clearly defined by the configuration of test vectors while others were indicated by the configuration but less clearly defined. All of these factors have been studied in subsequent test batteries in which each primary factor has been represented by new tests specially designed to feature the primary factors in the purest possible form. The object has been to construct tests in which there is a heavy saturation of a primary factor and in which other factors are minimized. This is the purification of tests by reducing their complexity.

These latter studies of the separate abilities were in each case made in the Chicago high schools—one study emphasizing the perceptual factor at the Lane Technical High School, one study of the inductive factor at the Hyde Park High School, an intensive study of the memory factor or factors in four high schools, and a study of numerical ability by Coombs in six high schools. In each series of tests, one factor was represented by a large number of tests, but all factors were well represented. In all of these studies the same primary abilities were identified as had been found in the experiment with college students. These studies led to the publication by the American Council on Education of an experimental battery of tests for the primary mental abilities, adaptable for use with students of high school or college age.

PRIMARY MENTAL ABILITIES OF CHILDREN

The identification of the same primary mental abilities among high school students as we had previously found among college students encouraged us to look for differentiation among the abilities of younger children. In the Chicago Public Schools, group mental tests are made of all 1B, 4B, and 8B children in the elementary schools and of 10B students in the high schools. The demand for a series of tests to be used in the guidance program for high school entrants and the advisability of not making too broad a leap in age led us to select an eighth-grade population for the next study.

The Eighth-Grade Experiment

In view of the purpose of investigating whether or not primary mental abilities could be isolated for children at the fourteen-year age level, the construction of the tests consisted essentially in the adaptation for the younger children of tests previously used with high school students. In some of the tests little or no alteration was necessary, while for other tests it was considered advisable to revise vocabulary and other aspects of the tests to suit the younger age level. A number of new tests were added to those selected from previous experimental batteries. Sixty tests constituted the final battery.

When the tests had been designed and printed, they were given in a trial form to children in grades 7A and 8A in several schools. Groups of from 50 to 100 children in these two grades were used for the purpose of standardizing procedures and, especially, for setting time limits.

Fifteen Chicago elementary schools were selected by Miss Minnie L. Fallon, Assistant Superintendent in charge of elementary education, and by Dr. Grace E. Munson, Director of the Bureau of Child Study, as experimental schools for this study. The tests in the main investigation were administered in the schools by the adjustment teachers. These adjustment teachers had had special training in testing procedures with the Bureau of Child Study and also had had considerable

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

experience in giving psychological and educational tests. Special instructions in the procedures for these tests were given to the adjustment teachers, as well as written instructions for each day's testing program.

Eleven hundred and fifty-four children participated in this study. The complete battery of 60 tests was given in 11 one-hour sessions to the children in the 8B grades in each of the 15 schools. The children enjoyed the tests and, with very few exceptions, the sustained interest and effort were quite evident. One thing which a psychologist might fear in such a long series of tests would be fluctuating motivation on the part of the students. Although the adjustment teachers administered the tests, every session was observed by a member of our staff, and we were highly gratified by the sustained interest and effort of the pupils.

In addition to the 60 tests we used three more variables: chronological age, mental age, and sex. The latter test data were available in school records. They were determined by the Kuhlmann-Anderson tests which had been given previously to the same children. Therefore, the battery to be analyzed factorially contained 63 variables.

The total population in this study consisted of 1,154 eighth-grade children. When all the records had been assembled, it was found that 710 of these subjects had complete records for all of the 63 variables. We decided to base our correlations on this population of complete records rather than to use the large population with varying numbers of cases for the correlation coefficients. For convenience of handling with the tabulating-machine methods, the raw scores were transmuted into single digit scores from which the Pearson product-moment correlation coefficients were computed. With 63 variables there were 1,935 Pearson correlation coefficients.

This table of intercorrelations was factored to 10 factors by the centroid method on the tabulating machines by means of punched cards. Successive rotations made by the method

PRIMARY MENTAL ABILITIES OF CHILDREN

of extended vectors yielded an oblique factorial matrix which is a simple structure

Inspection of the rotated factorial matrix showed seven of the factors previously indicated: Memory, Induction, Verbal Comprehension, Word Fluency, Number, Space, Perceptual Speed, and three less easily identifiable factors. One of these is another Verbal factor, one is involved in ability to solve pencil mazes, and one is present in the three dot-counting tests which were used

We have computed the intercorrelations between the 10 primary factors. Our main interest centers on the seven primary factors that can be given interpretation and, especially, on the first six of these factors for which the interpretation is rather more definite. Among the high correlations we note that the Number factor is correlated with the two Verbal factors. The Word Fluency factor has high correlation with the Verbal Comprehension factor and with Induction. The Rote Memory factor seems to be independent of the other factors. These correlations are higher than the correlations between primary factors for adults.

Because of the psychological interest in the correlations of the primary mental abilities, we have made a separate analysis of the correlations for those factors which seem to have reasonably certain interpretation. If these six primary mental abilities are correlated because of some general intellectual factor, then the rank of the correlation matrix should be one. Upon examination, this actually proves to be the case. A single factor accounts for most of the correlations between the primary factors.

The single factor loadings show that the inductive factor has the highest loading and the Rote Memory factor the lowest loading on the common general factor in the primary abilities. This general factor is what we have called a second-order general factor. It makes its appearance not as a separate factor, but as a factor inherent in the primaries and their correla-

tions. If further studies of the primary mental abilities of children should reveal this general factor, it may sustain Spearman's contention that there exists a general intellectual factor. Instead of depending on the averages or centroids of arbitrary test batteries for its determination, the present method should enable us to identify it uniquely.

We have not been able to find in these data a general factor that is distinct from the primary factors, but the second-order general factor should be of as much psychological interest as the more frequently postulated, independent general factor of Spearman. It would be our judgment that the second-order general factor found here is probably the general factor which Spearman has so long defended, but we cannot say whether he would accept the present findings as sustaining his contentions about the general factor. We have not found any occasion to debate the existence of a general intellectual factor. The factorial methods we have been using are adequate for finding such a factor, either as a factor independent of the primaries or as a factor operating through correlated primaries. We have reported on primary mental abilities in adults, which seem to show only low positive correlations except for the two verbal factors. In the present study we have found higher correlations among the primary factors for eighth-grade children. It is now an interesting question to determine whether the correlations among primary abilities of still younger children will reveal, perhaps even more strongly, a second-order general factor.

Interpretation of Factors

The analysis of this battery of 60 tests revealed essentially the same set of primary factors which had been found in previous factorial studies. Six of the factors seemed to have sufficient stability for the several age levels that have been investigated to justify an extension of the tests for these factors into practical test work in the schools. In making this extension we have been obliged to consider carefully the difference between research on the nature of the primary fac-

PRIMARY MINIMAL ABILITIES OF CHILDREN

tois and the construction of tests for practical use. Several of the primary factors are not yet sufficiently clear as regards psychological interpretation to justify an attempt to appraise them generally among school children. The primary factors that do seem to be clear enough for such purposes are the following: Verbal Comprehension V, Word Fluency W, Number N, Space S, Rote Memory M, and Induction or Reasoning R. The factors which in several studies are not yet sufficiently clear for general application are the Perceptual factor P and the Deductive factor D.

The Verbal factor V is found in tests involving verbal comprehension, for example, tests of vocabulary, opposites and synonyms, completion tests, and various reading comprehension tests.

The Word Fluency factor W is involved whenever the subject is asked to think of isolated words at a rapid rate. It is for this reason that we have called the factor a Word Fluency factor. It can be expected in such tests as anagrams, rhyming, and producing words with a given initial letter, prefix, or suffix.

The Space factor S is involved in any task in which the subject manipulates an object imaginally in two or three dimensions. The ability is involved in many mechanical tasks and in the understanding of mechanical drawings. Such material cannot be used conveniently in testing situations, so we have used a large number of tasks which are psychologically similar, such as Flags, Cards, and Figures.

The Number factor N is involved in the ability to do numerical calculations rapidly and accurately. It is not dependent upon the reasoning factors in problem-solving, but seems to be restricted to the simpler processes, such as addition and multiplication.

A Memory factor M has been clearly present in all test batteries. The tests for memory which are now being used depend upon the ability to memorize quickly. It is quite

possible that the Memory factor will be broken down into more specific factors

The Reasoning factor R is involved in tasks that require the subject to discover a rule or principle covering the material of the test. The Letter Series and Letter Grouping tests are good examples of the task. In all these experimental studies two separate Reasoning factors have been indicated. They are perhaps Induction and Deduction, but we have not succeeded in constructing pure tests of either factor. The tests which we are now using are more heavily saturated with the Inductive factor, but for the present we are simply calling the ability R, Reasoning.

In presenting for general use a differential psychological examination which appraises the mental endowment of children, it should not be assumed that there is anything final about six primary factors. No one knows how many primary mental abilities there may be. It is hoped that future factorial studies will reveal many other important primary abilities so that the mental profiles of students may eventually be adequate for appraising educational and vocational potentialities. In such a program the present studies are only a starting point in substituting for the description of mental endowment by a single intelligence index the description of mental endowment by a profile of fundamental traits.

The Final Test Battery

In adapting the tests for practical use in the schools for the appraisal of six primary mental abilities, we must recognize that the new test program has for its object the production of a profile for each child, as distinguished from the description of a child's mental endowment in terms of a single intelligence index. For many educational purposes it is still of value to appraise a child's mental endowment roughly by a single measure, but the composite nature of such single indices must be recognized.

PRIMARY MENIAL ABILITIES OF CHILDREN

The factorial matrix of the battery of sixty tests was inspected to find the three best tests for each of seven primary factors. In making the selection of tests for each primary factor we considered not only the factorial saturations of the tests, which are, of course, the most important consideration, but also the availability of parallel forms which may be needed in case the tests should come into general use. Ease of administration and ease in understanding of the instructions are also important considerations.

The three tests for each primary factor were printed in a separate booklet and the material was so arranged that the three tests for any factor could be given easily within a 40-minute school period. The main purpose of the larger test battery was to determine whether or not the primary factors could be found for eighth-grade children, but the purpose of the present shorter battery was to produce a practical, useful test battery and to check its factorial composition. The selected tests were edited and revised so that they could be used for either hand-scoring or machine-scoring. The Word Fluency tests constitute an exception in that none of the tests now known to be saturated with this factor seems to be suitable for machine-scoring.

In order to check the factorial analysis at the present age level, we arranged to give the selected list of 21 tests to a second population of eighth-grade children. The resulting data were factored independently of the larger battery of tests. There were 437 subjects in this population who took all of the 21 tests. This population was used for a new factor analysis. The results of this analysis clearly confirmed the previous study. The simple structure in the present battery is sharp, with only one primary factor conspicuously present in each test, so that the structure could be determined by inspection for clusters.

A battery of 17 tests has been assembled into a series of test booklets for use in the Chicago schools. An experimental edition of 25,000 copies has been printed, and the plan for

securing norms on these tests includes their administration to 1,000 children at each half-year grade level from grade 5B through the senior year in high school. These records have been obtained during the school year 1940 to 1941. The use of such a wide age range in standardizing the test is at first thought, perhaps, rather strange. The effort was made in order to secure age norms throughout the entire range of abilities found among eighth-grade children since the tests are to become a part of the testing procedure for all 8B children in the Chicago schools. Separate age norms will be derived for each of the six primary abilities. If a single index of a student's mental ability is desired, it is recommended that the average of his six ability scores be used.

As soon as the norms are established, the tests will be published by the American Council on Education under the title "Chicago Primary Mental Abilities Tests." It is expected that the tests will be ready for distribution during the summer of 1941. The norms provided with the tests will be of a wide enough range to make the tests useful at the high school and upper grade levels.

The complete test program consists of 17 tests, all of which have been reduced to machine-scoring form except the three tests for the Word Fluency factor W. In the nature of the case there seem to be difficulties in reducing this test to machine-scoring form, and hence it has been retained in hand-scoring form. It should be said, however, that the W tests can be scored almost as fast, if not as fast, as the tests which are machine-scored. Since all of the tests can be hand-scored, their use is not limited to schools large enough to avail themselves of the scoring machine. The hand-scoring of all the tests is very easily accomplished by the use of perforated stencils to be provided with the tests. Hand-scoring is facilitated by the use of the scoring board distributed by the Stoelting Company.

The new battery represents six primary mental abilities, namely, Verbal Comprehension V, Space S, Number N, Mem-

PRIMARY MENTAL ABILITIES OF CHILDREN

ory M, Word Fluency W, and Reasoning R. They enable the skilled psychologist to tabulate a profile of six linearly independent scores instead of a single measure, such as the intelligence quotient.

Principals, teachers, adjustment teachers, and school psychologists have expressed their satisfaction with the profile of abilities plotted for each child. Probably the children themselves have found the profiles most interesting and have profited most from an examination of their own profiles. In the school year 1941-1942, these tests will be installed as a part of the educational guidance program in the Chicago schools by administering them regularly to 8B elementary school pupils and 10B high school pupils.

Some of the features of the tests should be mentioned. The tests are so arranged that machine-scoring and hand-scoring tests are directly comparable and will have the same norms. The child's task does not vary with the type of scoring, only the scorer's job is changed. Another feature is the use of fore-exercise booklets printed on yellow paper. The time limits for the practice exercises are approximate. When a test proper is started, the student places his white test booklet on top of his yellow practice booklet, and the examiners and proctors can check at a glance that every child is working in the right place. The tests proper are to be timed exactly. The three tests of each of the six abilities are arranged in a booklet for administration within a 40-minute school period. It is recommended that the successive booklets be given on successive school days.

Further Problems

One of our principal research interests at the present time is to determine whether primary abilities can be identified in children of kindergarten or first-grade age. A series of about 50 tests is well under way, and some of them are now being tried with young children. If we succeed in isolating primary abilities among these young children, our next step will be to prepare a practical battery of tests for that

age. A subsequent problem will be to make experimental studies of paper-and-pencil tests for appraising the primary abilities of children in the intermediate grades, approximately at the fourth-grade level. We are fairly confident that such tests can be prepared for use in the intermediate grades.

It is a long way in the future, but it is interesting to speculate on the possibility of using the tests of the primary mental abilities as the tool with which to study fundamental psychological problems of mental growth and mental inheritance. Absolute scaling of the tests at the different age levels will make possible studies on the rates of development of the separate abilities at various age levels. Modifiability of the abilities will be another problem to which we shall later turn attention.

A STATISTICAL EVALUATION OF CLINICAL COUNSELING¹

F. G. WILLIAMSON AND L. S. BORDIN

University of Minnesota

SYSTEMATIC efforts at evaluation are a relatively recent development in the field of counseling. The form of appraisal has ranged from "verbal research" to simple statistical analysis. Not all of these studies have avoided the pitfalls, of which there are many, to be found in this undertaking. The assumptions, methods, and weaknesses involved in the various evaluation approaches are summarized in a previous paper (10).

The present paper, one of a number to be reported, summarizes an experiment designed to evaluate a certain type of counseling. Since our conclusions are applicable only to counseling based upon the philosophy and procedures employed at the Testing Bureau of the University of Minnesota, this type of clinical counseling should be described.

This clinical counseling has as its purpose assisting the student to choose and make progress toward educational and vocational objectives which will yield maximum satisfaction. It is assumed that this end can be accomplished best by aiding him to set his aspirations in terms of the level of his potentialities. Naturally his potentialities must first be analyzed before a diagnosis of any discrepancy between aspiration and ability can be made and before assistance can be forthcoming from the counselor.

¹Assistance in the preparation of this material was furnished by the personnel of Work Projects Administration, Official Project No. 65-1-71-140, Sub-Project No. 93.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

The case data upon which analyses and diagnoses are made consist of standardized ability tests, personality and interest inventories, questionnaire records, and non-quantified information collected from the student, his associates, and his parents. This accumulation of information must be interpreted on the basis of an integrated picture of the individual provided through personal interviews. In other words the counselor deals with a unique individual rather than with a generalized conception of a group of individuals.

The interview provides the medium through which counseling is personalized and through which the student is assisted in making his decisions. While the decisions that the student accepts are and should be his own, the counselor sometimes plays a persuasive role in that he organizes relevant case data to highlight the alternative courses of action from which the student chooses. Once the student has made the choice, the counselor has the task of aiding him to orient himself to his interests, attitudes, and abilities, and his environment, home and family, recreation, and education for the most successful achievement of the chosen objective.

A fuller description of this clinical counseling process has been presented elsewhere (9, 11). Only by means of an accurate conception of what the counselor is doing and what he is trying to do can any evaluation of that counseling be meaningful. Moreover, we should not attempt to generalize our conclusions to include any other type of counseling than the one studied.

In attempting to evaluate this clinical counseling, we believe that a criterion flexible enough to avoid artificial fragmentation of the individual provides the most adequate design for experimentation. Essentially such a design involves a judgmental comparison of the individual's adjustment status before and after counseling. This method—essentially the non-statistical weighting of variables to form a composite estimate—was used in a previously reported study (11, chap. IX). In this study an estimate of the degree of the student's cooperation was used as a means of control. The control

EVALUATION OF CLINICAL COUNSELING

lies in the comparison of those students who did with those who did not follow the counselor's recommendations

The process of making these evaluative judgments involves three phases: (a) the preliminary review or analysis of the case data, (b) the follow-up interview, (c) the case evaluation

In the first phase of the experiment, all student cases were independently and critically read by two trained workers whose functions were to analyze and record all information contained in the case folder. Any discrepancies between the analyses of the two readers were reconciled or adjusted in conference with the staff members concerned with the project. The case reviewers also compiled questions concerning the present status of the student, his adjustment to his original problem or problems, his adjustment to the counsel given, and any other pertinent information. These questions were used subsequently in the follow-up interview.

For the second step all student cases were called in for a follow-up interview. Cases which were incomplete because of insufficient interview contacts or incomplete test battery and which could not be reached for follow-up interviewing on the campus were reached either through a questionnaire or an interview in the home. For those students who had left the University, information was collected, and used, concerning their adjustment to their jobs and their satisfaction with that out-of-school adjustment. The follow-up interview yielded information concerning the extent of success or failure achieved by the student in solving each of his original problems and the extent to which the counselor's advice had been followed subsequent to the original counseling interviews. The student's own statement of the degree of his satisfaction with his solution of the problems and with University Testing Bureau counseling and recommendations or any other interpretations or evaluations that he made were specifically recorded. The interviewer did not interpret or evaluate this information. The purpose of the follow-up interview was, essentially, to obtain the factual data on the present status of the case

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Trained evaluators next critically reviewed the original case data and the follow-up interview report to arrive at a judgment of the extent to which the student had adjusted the problems for which he had originally sought counseling. The effectiveness of the counseling was evaluated in terms of the following counseling functions or services.

1. Diagnosis of the student's vocational and educational possibilities

2. Advice in making appropriate choice of a vocational field and in securing the related educational training sequences.

3. Counseling as to recognition of, and alleviation of, disturbing factors (emotional, educational, economic, health) which may interfere with or prevent the acceptance of proper vocational choice and the achievement of appropriate training

4. Assistance in the discovery and utilization of personal resources in effecting an adjustment

5. Guidance in the use of all University personnel resources in diagnosis and counseling

The student's adjustment with regard to vocational choice and his progress toward achieving satisfactory training for that choice were judged by means of the following criteria

1. Choices made in line with aptitudes, interests, work habits, personality, etc

2. Program of studies in line with these choices and the student's qualifications.

3. The student's satisfaction with vocational choice.

4. Progress in achieving training for objectives in terms of the capacity of the student to profit from such training

5. Alleviation of factors which interfered with the making of a satisfactory vocational choice and with acquiring the necessary training, e. g., parental dominance of choice, inadequate study skills, etc

In making their appraisals the evaluators studied the student's interests and aptitudes, the counselor's interview notes, the student's reported comments, and his grade record achieved before and after counseling. All of the information was weighed and balanced with reference to the five criteria

EVALUATION OF CLINICAL COUNSELING

before a judgment was made of the degree of adjustment achieved by the student subsequent to counseling. The degree of cooperation was independently judged in the same manner.

The following five categories² formed the scale of adjustment.

Satisfactory Adjustment—1. The student is satisfied with his vocational adjustment at the time of the follow-up interview. In some cases the student's dissatisfaction will not be a deterrent to a rating of satisfactory adjustment. In instances where the student's aspirations are far above his level of ability, he is considered satisfactorily adjusted if he accepts the fact that his ambitions must be pitched at a lower level.

2. In the interviewer's judgment the vocational choice and adjustment of the student are adequate, based upon aptitudes, interests, and subjective factors revealed through interviews.

3. There has been an alleviation of distracting factors which interfere with vocational choice and professional training such as inadequate socialization, mental conflicts, financial problems, health handicaps, and any other problems.

4. Achievement in a given training program is commensurate with aptitudes and interests.

Some Progress Toward Adjustment—The student has not yet reached a satisfactory adjustment, according to the previously stated criteria, but is evidently started on the road and may eventually reach the desired objective. He may have come to the counselor with a number of problems involving vocational choice, classification in college classes, and social adjustment and personal peculiarities. In the follow-up interview it may be found, for example, that he has succeeded in settling his vocational question but that he is still struggling for mastery in regard to social adjustments.

No Change—This classification is used for those cases in which the problems remain the same as at the time of counseling. While the passage of time will usually make a problem more serious, the designation of "slightly worse" was not ap-

²Described and illustrated in an earlier study. See reference 11, chap. IX.

plied unless a choice point had actually been passed. Thus a sophomore who has not yet made a vocational choice would be classed as unchanged, but a junior without a vocational choice would be in a more serious position and therefore would be classified as slightly worse. Juniors should have begun specialization if they are to make "normal" progress toward graduation.

Slightly Worse—This is a condition in which the solution of the original problems seems slightly more remote and the factors which existed at the time of the first counseling contact still exist and are accentuated.

Much Worse—Those cases where the student's problems are more severe and the solution much more remote or less probable of achievement.

The judgments of the degree of cooperation were based upon the following categories.

Followed advice wholly—The student followed the counselor's advice with respect to the most dominant or important original problems.

Followed advice in part—The student either partially followed the counselor's advice with respect to the chief problems or completely followed advice with respect to some but did not follow advice with respect to others.

Did not follow advice—The student did not follow the counselor's advice in regard to any of the main problems.

In order to determine the reliability of classification of cases according to the foregoing two sets of categories, an "outside" judge was called in to make independent judgments of the adjustment of a random sample of 247 cases. A coefficient of correlation of .82 was found between the "outside" judge's classifications and those made by the evaluators. This coefficient may be interpreted as a high index of validity or as a fair index of reliability, according to the reader's own conception of the meaning of these two terms. In over half of the cases where a discrepancy occurred, the "outside" judge had made a higher classification than had the two original judges. Thus

EVALUATION OF CLINICAL COUNSELING

would seem to indicate that the evaluators had not over-estimated the effectiveness of the counseling

The question arises as to how much influence the student's subsequent academic achievement (available to the evaluators) had on the judge's estimate of adjustment and cooperation. The correlations between honor-point ratio achieved after counseling and judgment of adjustment were .23 and .39 respectively, for General College and the College of Science, Literature and the Arts³. The difference between these coefficients, of borderline significance ($D_r/S.E. D_r = 2.0^1$),

may indicate that academic adjustment is more closely related to judgment of total adjustment for SLA students than for General College students. This does not seem unreasonable, since SLA students are generally committed to careers in which academic achievement is one of the most immediate requisites for success. The correlations of honor-point ratio with judgment of cooperation were of negligible magnitude $-.16 \pm .05$ for General College students and $-.17 \pm .03$ for SLA students.

In all, data were collected on 987 complete student cases who used University Testing Bureau services during the years 1933-34-35. For the purposes of this study it was deemed desirable to analyze as homogeneous a population as possible without the sacrifice of too much data. For this reason 498 students from SLA and 195 students from the General College were selected. Classified according to their status at the time of counseling, in the SLA group were 154 pre-college cases, 176 freshmen, and 168 sophomores. The General College group contained 41 pre-college cases, 125 freshmen, and 29 sophomores. The pre-college cases were high school seniors or recent graduates who came to the Bureau for counseling in the spring and summer immediately preceding enrollment in the University.

That the groups chosen for study were a fairly satisfactory representation of the total range of ability and achievement in the undergraduate classes of these colleges can be demon-

³Hereafter designated as SLA

¹Computed according to Fisher's method (1 pp. 208-10)

stated from the distributions of aptitude test and high school percentile scores. Of the SLA students, 62.6 per cent and 78.1 per cent fell at or above the fiftieth percentile in aptitude and high school percentile scores, respectively, while 82.8 per cent and 65.2 per cent of the General College students fell at or below the fiftieth percentile in the same variables. Too often there is the tendency to assume that only low ability students have a desire for or need of counseling. The distributions of the SLA group would seem to refute this assumption. The General College population provides us with the opportunity to determine whether counseling can be equally effective with low ability students.

The representativeness of our groups in terms of SLA and General College freshmen was determined by a comparison of high-school average grades transmuted into percentiles. Unfortunately, statistics on sophomores in these colleges were not available. Because of the known elimination of freshmen with lower percentiles, the sophomore population should be higher on the average. Since our experimental population consisted of students from both classes, this analysis of representativeness is not precise. The freshmen in our group were compared with representative SLA and General College samples. For SLA the combined mean for a sample of 2,157 freshmen students of the fall classes of 1933-34-35 was 65.45 as compared with 69.59 for our experimental freshman group. Although this small difference of 4.14 is reliable ($C.R.$ of 3.39), it does not represent a very significant one as far as the purpose of this study is concerned. The comparison of the means of the General College groups yields similar results. The combined mean of representative freshmen of the 1933 and 1935 freshman classes was 34.00⁶, for our group it was 40.25. Although the difference is somewhat larger than that in the SLA group, it is not so reliable ($C.R.$ of 2.64).

We may conclude from these two analyses of the nature of our counseled groups that they were generally representa-

⁶From unpublished data collected by Dr. Ruth Eckert of the University of Minnesota.

EVALUATION OF CLINICAL COUNSELING

tive of their total populations and of the total range of ability to do college work. It is interesting to note that the students who are counseled by the Testing Bureau, contrary to the opinions of many, are not the students of inferior ability, but, if anything, are slightly superior to the general undergraduate population of these two colleges.

Results of the Experiment

Degree of Cooperation and Adjustment—Previous studies of the effectiveness of counseling which used similar methods have reported results in terms of percentages. The proportions of our groups who were classified as satisfactorily or partially adjusted (82.8 per cent of SLA and 86.2 per cent of General College) compare favorably with those reported in English studies. Oakley (3) and Macrae (2), working with small populations of younger students, reported 95 per cent and 55 per cent, respectively, as the proportions who followed advice and who were satisfied and successful in their occupational adjustment. Rodger (4), with a larger population, reports 79 per cent successful adjustment. Seipp (5), using a methodology almost identical with ours, analyzed the case records of 100 adults diagnosed and advised by the Adjustment Service of New York. She found that 57 per cent made a satisfactory adjustment subsequent to counseling. Our results are even more impressive when analyzed in terms of those who cooperated in following the counselor's advice. In these terms the percentage of the SLA students satisfactorily or partially adjusted is 93.5 and the percentage of General College students is 96.3.

Our data also indicate that the counseling was equally effective, if not more so, in gaining the cooperation of the student. For SLA 70.9 per cent cooperated wholly and 20.1 per cent partly, while for General College the percentages were 69.7 who cooperated wholly, and 24.1 partly. Viteles (7), diagnosing and advising 75 adolescents, found that 58 per cent followed advice completely and 21 per cent partly.

Since the SLA and General College groups differed so markedly in college aptitude, it was interesting to determine

whether there was any real difference between the adjustment classifications of the two groups of experimental cases. To test this hypothesis, the chi-square test of independence was used.⁹ The result (chi-square value of 3.48, $p > .05$) indicates that there was no difference in the adjustment achieved by the two groups. A similar analysis of the cooperation classifications yields a similar result (chi-square value of 2.51, $p > .05$). A further analysis involving the length of the student's attendance in the University was generally unrelated to either adjustment or cooperation. The chi-square values for the groups were insignificant in value.

Adjustment versus Degree of Cooperation—In addition to these direct analyses, we have attempted to shed light on the definition of the conditions which make adjustment more probable as an outcome of clinical counseling. First and foremost of these conditions is that the student cooperate with his counselor. Anyone who has had intimate experience in counseling will have observed that the cultivation of a cooperative attitude usually precedes effective counseling. That the greater proportion of adjusted students found among those students who cooperated is not accidental is clearly indicated by the test results of the independence of these two variables. The chi-square values of 115.62 and 47.44 for SLA and General College students are both highly significant ($p < .01$). This means that we may assume that a student who cooperates with the counselor in attempting a solution

⁹This statistic may be used to test the independence of distributions from a real or hypothetical distribution (6 chap. 1). As used in this study, the expected distributions were based upon the proportions of the five classes of adjustment observed in the total distribution. The formula for computation appropriate for this type of analysis is

$$\sum_n \left[\frac{(f_o - f_e)^2}{f_e} \right],$$

yielding a value which, by use of a table of chi-square distributions, is translated into an estimate of the probability that such a value could have been obtained for additional samples drawn from the same general population. We shall use the conventional five per cent and one per cent points as our confidence limits. These points are equivalent to values two and three standard deviations from the mean. Because of the small number of cases, some of the categories were combined in all of the chi-square tests used. Five was the smallest number of cases permitted in any one category.

of problems will in all probability achieve satisfactory adjustment as defined above. Only the General College sophomores did not exhibit a statistically significant relationship. The restriction of the range of the variables necessitated by the small number of cases in this group may explain this fact.

Expectancy of Adjustment According to Type of Problem—A previous study by Williamson (8) has shown that counselors tend to specialize in the types of problems that they treat. It is important, therefore, to determine the effectiveness of counseling with respect to different types of problems. Since in most cases students experience more than one problem, classification in any one category, e.g., vocational problem, will include students who may also have an educational problem, a social problem, or any other problem or combination of problems. In view of this fact, if the vocational category shows a significantly greater proportion of adjusted students than the educational or the emotional category, then evidence of the differential effectiveness of the counseling will have been discovered.

An analysis of our data gives a clear indication that the adjustment expectancy is not so marked for social-personal-emotional problems as for vocational and educational types of problems. While not all of the differences are statistically significant—the total General College group and the General College freshmen showed the significant ones (chi-square, 29.84, $p < .01$ and chi-square, 32.99, $p < .01$)—the trends are consistently in the same direction. Since it has already been shown that cooperation can be assumed as a counseling condition necessary to adjustment, it is not surprising to find that there is a greater expectancy of cooperation for vocational and educational problems than for social-personal-emotional ones.

Expectancy of Adjustment According to Status of Vocational Choice—Since the counseling being evaluated in this experiment is primarily educational and vocational, the types of changes in vocational orientation required should be of importance for the expectancy of adjustment. Four possibili-

ties were defined (a) confirmation of the student's choice by the counselor, (b) recommendation by the counselor of some choice other than the student's, (c) recommendation of a choice by the counselor, because of the student's indecision at the time of the original contact, (d) deferment of choice on the counselor's advice at the time of original contact. It had generally been assumed that the counselor is more likely to bring about adjustment when he has only to confirm the student's previous choice. The results of our experiment do not support this assumption. They indicate that it makes no difference, for this type of counseling, whether the student's choice was confirmed or changed or whether he was undecided at the time of the first interview. But in those cases where choice is deferred, the expectancy of adjustment is significantly less (chi-square of 28.59, $p < .01$). In the case of cooperation, what "ought to be true" actually is true. As one would suppose, greater cooperation is to be expected from those students whose vocational choices were confirmed (chi-square of 15.7, $p < .01$).

Aptitude and Achievement in Relation to Adjustment and Cooperation—One might expect that ability and previous achievement of students who come for counseling would be positively related to expectancy of cooperation and adjustment. This problem was attacked by studying the aptitude and achievement characteristics of students in each of the cooperation and adjustment categories. The analysis of the variance in aptitude test scores gives evidence that this assumption cannot be held in terms of the ability test used.⁷ The variance ratios were of such a small degree that the probabilities that they represented the same population were greater than five in a hundred. This means that low ability students are just as likely to be cooperative and adjusted as high ability students.

On the other hand, the analysis gives reliable evidence that high school achievement is positively related to coopera-

⁷Snedecor's tables of F (6 p. 174) were used to estimate the probabilities of getting as large a variance ratio from samples of a homogeneous population. Here again the five per cent and one per cent points were taken as the limits of confidence.

EVALUATION OF CLINICAL COUNSELING

tion and adjustment (General College, F is 8.09, $p < .01$, SLA, F is 5.45, $p < .01$) This relation is further emphasized by the finding that for any degree of adjustment, those students who cooperated were, for the most part, those with previously higher achievement. Previous college achievement could be analyzed validly only in relation to cooperation, since it already had entered into the estimate of adjustment. The results here are not conclusive. While the SLA data yielded a significant variance ratio (8.22, $p < .01$), the General College data did not.

Number of Interviews versus Adjustment and Cooperation—With respect to SLA students, variation in number of interviews indicates that the counselor had the most interviews with students who were partially adjusted (General College, F is 4.13, $p < .05$, SLA, F is 20.84, $p < .01$). Thus those students who are satisfactorily adjusted, characteristically, do not require so many interviews. Likewise, those students whose maladjustment is of such a nature (e.g., very low ability) as to offer little probability of adjustment are not interviewed so frequently. It seems, then, that those students who present difficult problems but give promise of adjustment seek counseling interviews most frequently. For SLA there is slight evidence for a positive relationship of number of interviews with judgment of degree of cooperation (F is 3.07, $p < .05$). In the case of General College students, there is a negative relationship between adjustment and the number of interviews. The students in the low adjustment categories apparently show a greater willingness or are more encouraged to return for further counseling than are the more satisfactory adjustment groups.

Time Interval versus Evaluation—The significance of the time elapsed between the first counseling interview and the follow-up interview should be of value in indicating the optimum period for an evaluation experiment. Since there is no reason to suppose that special selection has operated in the selection of the time at which the adjustment groups were studied, it is reasonable to infer that observed differences are

differences in time necessary for reaching that level of adjustment. While the previous analysis of SLA data indicated a greater number of interviews for the students classified in the partially adjusted category, it is evident that students in this category required the shortest time to achieve their degree of adjustment. Analysis of the General College data supports this result, a perplexing one. However, a more interpretable result is secured when the data are analyzed in terms of the interrelation of cooperation and adjustment. The trend is in the direction of a shorter time interval for students in any degree of adjustment who cooperated to a greater degree with the counselor. The inference may be made that those students who cooperated reached a given level of adjustment in a shorter time. The difference averages a little over two months in an average evaluation period of 16 months. The F value of 3.4 is beyond the one per cent point.

Summary

This evaluation of the clinical counseling practiced in the Testing Bureau of the University of Minnesota has attacked two basic problems: (a) What proportions of students were aided by the Bureau's counseling to achieve a better adjustment? (b) What conditions and characteristics of counseled students are most conducive to a favorable prognosis of subsequent counseling?

In answer to the first question, counseling was effective in achieving the cooperation of and in improving the adjustment of over 80 per cent of the students in our groups. This is especially significant in that the analysis and classification of cases were carefully defined and controlled, having been made by judges who had not been involved in any of the counseling.

The conditions and characteristics favorable to adjustment include the following:

1. Cooperation with the counselor was positively related to adjustment and those students who cooperated reached their level of adjustment in a shorter period of time than those who did not.

EVALUATION OF CLINICAL COUNSELING

2 Students experiencing educational and vocational problems were more successfully counseled than were those with dominant social-personal-emotional problems

3 Contrary to belief, our data indicate no differences in adjustment among counseling cases classified as vocational choice confirmed, altered, or undecided at the first contact. But, if vocational choice is deferred by the counselor, the prognosis of adjustment is less favorable

4 Higher high school or previous college achievement is positively related to cooperation and adjustment. But level of ability, as measured by the aptitude test used in this experiment, is not related

These conclusions may be interpreted as limitations either of the students involved or of this type of counseling. In the case of the type of problem, it is likely that a limitation of counseling is disclosed. Counseling that is educationally and vocationally oriented is not likely to deal so effectively with social-personal-emotional problems. On the other hand, it does not seem probable that any type of counseling or improvement in treatment techniques can do much for a student with a very low achievement background insofar as the types of adjustment involved in this evaluation experiment are concerned.

Certain relations of an ambiguous nature and therefore demanding further study were observed. There was evidence that the counselor conducts more interviews with students who are judged as partially adjusted, yet this same group reached their level of adjustment within a shorter period of time. Our data do not indicate whether or not the counselor tends to intensify his work with certain students by conducting many interviews within a short period of time.

There is one conclusion that this study should have made clear. The evaluation of counseling is not a casual process, easily carried out. Indeed, such a study represents a combination of careful and rigorous case reading, many days and weeks of interviewing, prolonged clerical and statistical labor, and above all a period of patient waiting for the counseling cases

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

to mature to the stage wherein adequate data are available for critical evaluation

REFERENCES

- 1 Fisher, R. A. *Statistical Methods for Research Workers* (7th ed.) London: Oliver and Boyd, 1938. 356 pages
- 2 MacRae, A. "A Follow-up of Vocationally Advised Cases," *Journal of the National Institute of Industrial Psychology*, V (1931), 242-47
- 3 Oakley, C. A. "A First Follow-up of Scottish Vocationally Advised Cases," *Human Factors* (London), XI (1937), 27-31.
- 4 Rodger, T. A. "A Follow-up of Vocationally Advised Cases," *Human Factors* (London), XI (1937), 16-26
- 5 Seipp, Emma. *A Study of One Hundred Clients of the Adjustment Service* (Adjustment Service Series, Report XI). New York: American Association for Adult Education, 1935. 30 pages
- 6 Snedecor, G. W. *Statistical Methods*. Ames: Collegiate Press, Inc., 1937. 341 pages
- 7 Viteles, M. S. "Validating the Clinical Method in Vocational Guidance," *Psychological Clinic*, XVIII (1929), 69-77
- 8 Williamson, E. G. "Faculty Counseling at Minnesota: An Evaluation Study of Social Case Work Methods," *Occupations*, XIV (1936), 426-33
- 9 Williamson, E. G. *How to Counsel Students*. New York: McGraw-Hill, 1939. 561 pages
- 10 Williamson, E. G. and Bordin, E. S. "Evaluation of Vocational and Educational Counseling: A Critique of the Methodology of Experiments," *Educational and Psychological Measurement*, I (1941), 5-24.
- 11 Williamson, E. G. and Darley, J. G. *Student Personnel Work*. New York: McGraw-Hill, 1937. 313 pages

CONTRIBUTION OF TESTS TO RESEARCH IN THE FIELD OF STUDENT PERSONNEL WORK

RALPH W TYLER

University of Chicago

THE USE of tests is fundamental to many aspects of student personnel work. In the selection of students, in identifying their potentialities, their problems, and their difficulties, in checking on the effectiveness of procedures used in providing for personal development, in vocational placement and follow-up, personnel workers have learned to use a wide range of tests and to depend on the results of tests as a basic part of the personnel program. Although the place of tests in the practice of personnel work has been well outlined, the contribution to be expected from tests in connection with personnel research has not been so clearly indicated. I am differentiating research from practice in the field of student personnel work by defining research as the process by which basic facts, theories, principles, instruments, and procedures are developed, thus providing a rational framework upon which the practice of student personnel work can be understood and elaborated. This distinction may perhaps be made clearer by illustration.

It is a common *practice* of the student personnel officer to administer reading tests to incoming freshmen, to study the results, to identify certain students who received relatively low scores on the reading tests, and to recommend a remedial program in reading for some of these students. *Research* which finds out what reading demands are likely to be made by the various freshmen courses, which devises valid instru-

ments for measuring these reading abilities, which estimates the probable frequency of inadequate reading abilities among freshmen, which develops theory and principles regarding the relation of reading development to other aspects of the student's development, and which establishes the probable validity of various types of remedial reading procedures would represent the essential framework upon which improved personnel practices relating to reading can be built. Practice and research are complements in a sound professional growth.

Because student personnel work may be concerned with all aspects of the personal and social development of students, its problems relate to many previously organized fields of research such as physiology, psychology, sociology, anthropology, psychiatry, and education. Obviously, personnel workers have drawn and must draw upon these various organized fields of research for many of their concepts, instruments, and practices. However, many problems which the student personnel worker faces cut across two or more of these fields and are likely to involve research aspects not adequately investigated by any one of these disciplines alone. The problems which do involve two or more organized disciplines are the problems which in general must be attacked by research workers in the field of student personnel. May I indicate some of these problems and suggest contributions which tests have made or can make to research on these problems?

One major research problem is to delineate clearly desirable goals for a student personnel program. Accepting the general function of personnel work to be the facilitation of well-rounded personal and social development of students, it is evident that this function must be defined more clearly in the case of a given college or type of college so as to indicate the aspects of development to be promoted and the desired relation among these various aspects. This clearer picture of the phases of student development to be given attention and their relation is essential to the intelligent direction of a program aimed at facilitating well-rounded development of the individual student.

TESIS IN STUDENT PERSONNEL WORK

It is obvious that a profession should have its goals clearly and definitely in mind; it is not so obvious that the formulation of these goals for the field of student personnel is a research problem of considerable magnitude. The difficulty of the task is partly due to the complexity of human development. Well-rounded personal and social development includes physiological, psychological, and social aspects. Furthermore, these various aspects are interrelated, that is to say, physiological development influences and is influenced by psychological and social development. Correspondingly, psychological development influences and is influenced by physiological and social development, and social development influences and is influenced by physiological and psychological development. Hence, although research in the several established disciplines helps to identify characteristics of normal physiological growth, of psychological maturation, and of social development, special research of a co-ordinated or integrated nature is necessary to establish the desirable balance among these several aspects of student development.

A second factor which complicates the formulation of goals for this field is the relation of student personnel work to the rest of the college program. In order that a college have the most effective influence upon its students the various phases of the college program, curricular and extracurricular, need to have some underlying coherence, that is, they must be bound together by common purposes. The major purposes of a college are educational, and the acceptable goals of student personnel services also should be at least in harmony with the educational purposes of the institution, and preferably they should serve to promote these educational purposes. In the actual practice of student personnel work there is danger that we shall carry on activities day after day without carefully considering their relation to the primary aims of our institution. This may lead to a short sighted program in which immediate goals are attained without really promoting the ultimate goals of the college. It is possible, for example, to work out a plan of housing which provides for very quick

adjustments of the students to their classmates, and yet by the nature of the housing plan, cliques may be encouraged and the fundamental educational objective of learning to understand people with very different backgrounds and to enter sympathetically into the lives of persons very different from ourselves may be hindered rather than promoted. Or, a social counselor may feel that her job is well done when she has helped to increase the proportion of women students who have regular dates, whereas the ultimate educational objective is to get a broader understanding of human behavior including a sympathetic understanding of and adjustment to the opposite sex. Continued dates with the same individuals in many cases may retard the attainment of this objective rather than help it. It seems necessary, therefore, to formulate goals for student personnel work in such a way that they are closely related to the major educational objectives of the institution.

This implies that the student personnel worker in close collaboration with other members of the school or college staff will need to examine the various types of studies which suggest possible goals of student development. They will need to consider the investigations of the sociologist, the social anthropologist, the social psychologist, the economist, and the political scientist, to identify the demands which our culture makes upon young people and to understand the effect of cultural pressures upon the individual and his group. These studies of the social scientists represent an important component from which goals for student personnel work will be formulated.

But an examination of results of research in the social sciences is not enough. It is also necessary to examine studies of student health and investigations in the fields of physiology, nutrition, and psychiatry—for these help to clarify the concept of desirable biological development and also to indicate possible deficiencies which students may be helped to overcome.

A third component of research regarding goals for student personnel work is the field of values. Values need to be considered carefully not only as possible student goals but also

because values, individual and cultural, condition the student's development in many ways. The ideals which young people absorb from contact with the culture have a more potent influence upon student goals, student activities, and the satisfactions and disappointments of college life than is commonly realized. Any comprehensive formulation of goals for student personnel work needs to consider the values which the school or college may be expected to promote and the way in which school or college experiences may influence these values.

I have suggested several of the strands which need to be considered in delineating goals for student personnel work. It is obvious that the selection of goals to be given particular emphasis in a particular college depends upon several factors. One is the college's conception of the good life and its counterpart—the desirable person. This conception will represent not only specific items such as physical health, social concern, personal integrity, and the like, but it will also involve some idea of the relation of these various aspects. At this point it is very necessary for the student personnel worker to have a workable but comprehensive theory of personality structure and function, and of personality development. Because we do know that the human organism shows a considerable degree of unity in its reactions, because we do know that physical, social, and psychological aspects are interrelated, we realize that one cannot treat each aspect of a student's development in isolation from the others. Some theory as to how these aspects are related, how they function together, and how they may be developed together is essential to provide a rational basis for personnel work. If the student personnel worker together with other members of the school or college staff has identified more specifically the aspects of human development which the school or college seeks to promote, and if he has a comprehensive theory of personality development, it is possible to formulate clear yet comprehensive goals for his own work and to avoid treating a student as though he were a mechanical collection of specific reactions.

What contributions have tests made or can tests make to

this area of research? A test provides a controlled situation in which certain specified types of behavior may be studied and certain phases of this behavior may be measured. In the effort to formulate a coherent theory of personal and social development, various types of tests must be constructed so that students can react in ways which involve the relation of biological, social, and psychological phases of behavior. These tests may enable us to see more clearly how these phases of behavior are related. Furthermore, any college, after determining the tentative goals of its student personnel work, can employ tests to determine which of these goals are of primary importance to its students. Conscious attention need not be given in the college program to those points at which students are already developing satisfactorily. That is to say, tests contribute to this area of research both in developing a comprehensive set of goals and in identifying the goals which need major attention in a particular college at a particular time.

A second area of research in the field of student personnel work is the testing of the fundamental bases upon which a student personnel program is built. A well-founded plan of personnel services is a recent addition to the college campus. Most of the schemes have been based upon assumptions which have not been adequately tested. The principles of organization, of administration, of the selection of the staff, of the training of the faculty—all are in need of careful verification. These principles seem to the administration or faculty of the given institution to be sound, but in many cases they have been drawn from fields and experiences which are not strictly parallel to the field of student personnel, and it is likely that some of these principles are not appropriate as part of the foundation of the program of student personnel services. Research provides a check on the validity of the basic foundation of the personnel program. Such research involves comprehensive evaluation of an entire personnel program or of particular procedures.

A comprehensive evaluation provides evidence showing how far each of the important objectives or goals of student personnel work is being attained. Since these goals involve various aspects of student development, tests of various sorts are essential in order to find out the points at which students are developing adequately or the points at which development is unsatisfactory. For example, this research requires tests of physical development, of health, of personal-social adjustment, of attitudes, of interests, of skills, of information acquired, and the like. It also involves a periodic program of testing so as to estimate the progress being made by the students, and correspondingly, their rate and degree of development. Furthermore, an adequate research program provides a follow-up of students after they have been graduated from college and have gone out into life. These follow-ups should probably be made from five to ten years after graduation and should include the collection of data regarding those objectives which have most permanent significance. This probably would include evidence regarding intellectual interests, health practices and attitudes, marital adjustment, social-civic interests and activities, and maturity of aesthetic interests. Such a follow-up study provides an important type of data regarding the continuing development of students and, therefore, it is a significant phase of the evaluation of the personnel program.

The checking of the fundamental bases upon which the student personnel program is built is an area of research which has largely been dependent upon valid tests. The recent accelerated development of a wider range of tests has been accompanied by a corresponding increase in evaluative studies. Tests are making an important contribution to this area of research.

A third area of research in the field of student personnel work which involves tests is the construction and validation of instruments to facilitate the personnel program. Tests represent the major group of these instruments. Various tests have been constructed for use in selecting students likely to

benefit from a given college program. Much research is still needed in identifying important characteristics of young people which can be used as a basis for college selection and for planning programs of educational and vocational guidance. Thus far, these tests have largely consisted of measures of verbal facility, of numerical manipulation, and of the acquisition of information. Tests of higher intellectual skills, of interests, of personal-social adjustment, and of attitudes are just beginning to contribute markedly to the selection and guidance work.

Tests already developed have greatly facilitated the identification of students needing special attention, but many new instruments are also needed. Tests are widely available for identifying certain types of reading difficulty and certain types of subject-matter deficiency. New instruments are needed, however, to measure other types of psychological and social reaction which have fundamental significance for success in college and in life, and which should be identified early enough so that a program to facilitate development may be begun.

A similar condition exists with regard to tests useful in the vocational placement of students. Tests of some of the essential vocational skills have been of great value. Tests for identifying certain vocational interests are showing promise. However, some of the fundamental vocational attitudes, habits, and ways of thinking have not been clearly identified, nor have satisfactory tests for them yet been developed. The future contributions of tests of this type are likely to be large.

New tests are being constructed to help in evaluating personnel programs and procedures. Judgments of students and faculty have not only been supplemented by more careful case studies and observational records, but tests of attitudes, of interests, of habits and practices, of information and skills are becoming available for a more comprehensive evaluation. Additional tests are still needed, and many are in the process of construction.

I have attempted to suggest briefly the place of tests in three areas of research, namely, in delineating goals for student personnel work, in checking the fundamental bases

upon which personnel programs and procedures are developed, and in constructing essential instruments for personnel work. Tests have already made an important contribution to these three areas of research, but the future contributions should be far greater than those of the past. The limitations of the contributions of the past seem to me to have been due to several factors which now can be largely overcome.

In the first place, student personnel work originated largely from specific maladjustments within the traditional college program. Particular problems relating to the conduct and morals of students, their social life, or their housing led to the provision of special staff members to handle these difficulties. Only within recent years has there been wide recognition of the broad implications of student personnel work and of the need for some coherent philosophy and program. Naturally, tests used in the student personnel field frequently were taken over, as they were developed, for other purposes and used without consideration of the behavior patterns which these tests implied. It seems to me that we are now ready to formulate a coherent conception of personal and social adjustment and to examine possible tests in the light of our concept, discarding or modifying tests which do not appropriately fit this concept and developing new tests that are in harmony with it.

With this bit-by-bit accumulation of personnel responsibilities in the college program, it was natural that the tests used should largely be built upon a type of atomistic concept of human behavior, and that the test results should be summarized as single scores or as separate parts added together to form a total score. In recent years we have seen more clearly how to construct tests involving greater organization of behavior and how to summarize results in terms of descriptive scores relating various parts of a test, thus getting a more coherent picture of the student's response. This elimination of the single composite score is an important step in increasing the contribution tests make to the field of student personnel work.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

An additional reason for my belief that tests will make an increasing contribution to the field of student personnel work is the wide recognition of a broad definition of tests. No longer are tests conceived only as paper-and-pencil examinations. Tests are increasingly considered as controlled methods for obtaining a sample of a student's reactions under certain specified conditions. With this recognition that a test is a means of sampling certain aspects of human behavior, attention is now being focused upon clearer definitions of those aspects of human behavior which need to be sampled by means of tests. In educational testing twenty years ago, primary attention was given to sampling the content of textbooks which students were expected to remember and to sampling certain of the subject-matter skills, such as writing or numerical computation. It is now recognized that other aspects of behavior are important, such as the way in which the student attacks problems, the types of interests he is developing, the attitudes he has, his response to aesthetic experiences such as literature, music, and the arts.

With greater clarification of the nature of testing has come a better specification of the behavior to be tested. Twenty years ago a test in chemistry would be built by specifying the topics, that is, the content to be sampled. No conscious effort was made to specify the type of reaction the student might be expected to make to this content. Now we recognize that we must specify not only the content but also the kind of reaction expected of the student, the sort of situation in which such reaction can be expected and, if possible, the kind of purpose which a student would have when reacting. By specifying these four aspects of behavior we have a much clearer idea of what we are trying to test, and this increases the probability that we shall control the testing situation sufficiently to provide a satisfactory test.

GRADE AND AGE NORMS FOR THE MINNESOTA VOCATIONAL TEST FOR CLERICAL WORKERS¹

GWENDOLLEN G. SCHIFFDIER

University of Minnesota

PROGRESS in the applications of psychology, especially in the field of aptitude measurement, will be made by working intensively on the measuring instruments which we already have, rather than by adding to the large number of devices about which we have insufficient research to justify scientific application. In line with this belief we have investigated certain problems connected with the Minnesota Vocational Test for Clerical Workers. The portion of this research to be reported here deals with a normative study of this test.

The usefulness of the Minnesota Vocational Test for Clerical Workers has been seriously curtailed because of the fact that norms have been established only for adults in the general population and for employed adult clerical workers (6). This limitation is the reverse of the more usual and serious one where norms exist for school populations while no adequate norms exist for adults and for criterion groups. The problem of appropriate norms for tests is one of the most urgent ones which counselors face in applying measuring instruments in guidance programs where individual analysis and diagnosis is an indispensable first step. The Minnesota

¹The cooperation of many persons has been necessary for the completion of this study and the author wishes hereby to express her appreciation. Professor Donald G. Paterson directed the construction of this measuring instrument by Dr. Dorothy M. Andrew and has followed through with advice and helpful suggestions in subsequent research. Assistance in the preparation of some of the materials for this study was furnished by the personnel of Work Projects Administration, Official Project Number 665-71-3 69, Sub Project Number 229.

Vocational Test for Clerical Workers was standardized on adults, and excellent norms were developed and reported in the Bulletins of the Employment Stabilization Research Institute (5, 9) and in the test manual (6). The test with its norms for adults was used to advantage by the Adjustment Service in New York City, a community guidance agency for adults, and by other agencies concerned with the counseling of adults. As the test has become more widely adopted, however, it has been applied in many situations, especially to youth populations for which the significance of test scores was not known. Some workers have devised local norms which have reflected selective factors of sampling. Limitations in interpretation have necessarily accompanied limitations in the selection of the sample. What has been needed is a normative study of this test based upon a large sample of youth representative of the populations at the junior and senior high school levels. With such research it becomes possible to apply the Minnesota Vocational Test for Clerical Workers to the age range for which the test is most appropriate from the standpoint of educational and vocational guidance.

The Minnesota Vocational Test for Clerical Workers is composed of two subtests. Test I consists of 200 paired numbers varying in length from three to 12 digits, Test II consists of 200 paired names varying in length from seven to 16 letters. Slight changes had been made in half of the paired items and the subject is asked to compare the paired items as rapidly as possible, checking those pairs which are identical. He is allowed eight minutes for Test I and seven minutes for Test II. Scores are calculated on each of the two subtests using the "right minus wrong" formula. The administration of the test is described in the manual and other sources (6, 4, 9).

The reader interested in research evidence of the test's reliability and validity, and in information regarding adult norms and the relationship between test scores and other variables is referred to the references on page 156 and especially

NORMS FOR MINNESOTA CLERICAL TEST

to the monograph by Andrew and Paterson (5). The following paragraphs summarize the research very briefly.

Andrew (5) has presented evidence on reliability which indicates that the test yields sufficiently stable results for use with individuals.

Andrew (5) has also presented a considerable body of evidence which points to the test as a valuable technique in a clinical program of educational and vocational guidance or selection to eliminate persons not likely to succeed in clerical training or employment. The test results correlate highly with high school and college teachers' ratings of clerical aptitude—in fact, higher than does a test of general intelligence. They also are definitely related to achievement records in typing and to criteria of production on clerical jobs as well as to supervisors' ratings of proficiency on the job. The test appears to be measuring factors other than academic intelligence or clerical training and experience and it is better than other tests for differentiating clerical workers from persons in the general population.

The relationship between the two subtests is not sufficiently high to justify using one test alone or combining the two scores (5). Reading speed is not an important factor in the test.

The method of scoring is that of "right minus wrong." Despite certain criticisms of this technique (7), it can be upheld on logical bases (10).

Significant sex differences on test scores have been reported (5) for men and women in the general population but not for men and women employed in the same type of clerical positions.

The test author (1, 2, 3) has made an analysis of the test to determine the abilities which it is measuring and has concluded that Test I involves a numerical factor and Test II a verbal factor and that both are relatively unrelated to academic intelligence, ability to perceive spatial relationships, and dexterity with fingers and small tools.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

To secure norms which would be fairly representative of a cross-section of junior and senior high school pupils in the North Central Association of Secondary Schools, St. Paul, a midwestern city of one-quarter of a million population, was chosen. Approximately 4,000 pupils in grades eight through twelve were given the Minnesota Vocational Test for Clerical Workers. This does not represent the entire population for these grades. In order to guard against a possible selective sampling in the choice of schools, an attempt was made to select at each grade level schools representing the upper, middle, and lower socio-economic groups. One high school and one junior high school, judged to represent each of these three groups, were chosen. To guard further against securing a selective sampling within the schools, the pupils were tested in English classes, since English is a subject required of all irrespective of curriculum followed. Table 1 shows the number of pupils included in the norms, distributed by grade, sex, and school. Schools A and B represent the above-average socio-economic groups, schools C and D represent the average, and schools E and F were characterized by a large proportion of families in the lower socio-economic groups.

The testing procedure was standardized and adhered to throughout the program with testing done in the regular English classes. The administration of the test was that prescribed by the test author (4, 5, 6). Personal data items including identifying data, date of birth, grade, school, curriculum, and father's occupation were filled out by the pupils on the last page of the test folder (10) before the test itself was administered. Birth dates were checked against school records. Additional data, such as high school scholarship percentile rank and intelligence test scores,² were collected for certain pupils and recorded on the personal data sheet. All tests were rescored at least once.

An important prerequisite to the publication of norms on tests which are to be used widely is a careful description of the population on which the norms were based. Only in this way

²The Aptitude Index of the Van Wageningen Unit Scales of Aptitude, Forms E, D, or C.

NORMS FOR MINNESOTA CLERICAL TEST

TABLE 1
DISTRIBUTION OF CASES BY GRADE, SEX, AND SCHOOL

DISTRIBUTION OF CASES BY GRADE, SEX, AND SCHOOL													
Grade	Male				Female				Male & Female				
	B*	F	D	Total	B	F	D	Total	B	F	D	Total	
VIII	113	65	106	284	152	60	96	288	245	125	202	572	
IX	B	F	D	C	B	F	D	C	B	F	D	C	
	105	106	2	121	117	96	0	114	220	202	2	255	
X	A	E	C		A	E	C		A	E	C		
	153	95	124	372	165	100	151	416	518	195	275	788	
XI	A	E	C		A	E	C		A	E	C		
	140	145	96	381	162	131	134	427	502	276	250	808	
XII	A	E	C		A	E	C		A	E	C		
	192	247	100	539	206	255	97	538	598	482	197	1,077	
Total				1,908	Total				1,996	Grand Total of Cases			3,904

*Letters designate the different schools

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

are test consumers able to determine whether or not the norms are appropriate and applicable to their local populations. Before presenting the tables of norms we shall, therefore, describe the sample of the school population to which we administered the Minnesota Vocational Test for Clerical Workers, presenting statistics on intellectual characteristics, age-grade locations, and socio-economic levels.

Table 2 describes a large proportion³ of our sample at each grade level in terms of the central tendency and variability of intelligence test scores. This table also gives similar figures for the total St. Paul school population for these

TABLE 2

COMPARISON OF THE MEANS OF THE ST. PAUL SCHOOL POPULATION^b AND A LARGE PERCENTAGE OF OUR SAMPLE ON THE BASIS OF THE UNIT SCALS OF APTITUDE "APTITUDE INDEX"

Grade	Groups	N	Mean	S. D.	Diff.	
					S. D.	Diff.
VIII	St. Paul School Population	2,327	105.3567	11.7982	2.4190	3.4366
	Our Sample	514	107.7757	11.7785		
IX	St. Paul School Population	2,169	104.7833	13.7310	3.7096	5.9130
	Our Sample	564	108.4929	13.2920		
X	St. Paul School Population	2,299	104.9674	13.1065	1.1290	2.0610
	Our Sample	716	106.0964	12.6935		
XI	St. Paul School Population	1,850	106.8270	12.2660	.8814	1.6866
	Our Sample	758	106.7084	12.0660		
XII	St. Paul School Population	1,323	108.3117	11.2040	— .9977	1.4996
	Our Sample	1,015	107.3470	11.9260		

*These figures for the St. Paul school population were provided through the courtesy of Professor M. J. Van Wageningen of the University of Minnesota.

³Ninety three per cent of the total 3,904 cases are so described. No intelligence test score for the remainder could be located but there is no reason to suspect the operation of a selective factor here.

grades. There is no necessity that our sample should be strictly representative of the St. Paul population. That would have been required if we had desired to develop norms appropriate only to the St. Paul school population at a particular date. Our purpose has been to develop norms on a sample judged to be fairly typical of the school population in North Central secondary schools and then to describe that sample as adequately as possible.

It can be seen that our sample differs from the St. Paul school population by from less than one to less than four points on the average, depending upon the grade. These small differences are more significant statistically for grades eight and nine than for grades ten, eleven, and twelve, as can be seen from the ratios of differences to the standard deviations of those differences. Using this test of representativeness, then, we can say that our tenth, eleventh, and twelfth grade students are, on the average, more like the St. Paul school population from which they were drawn than our eighth and ninth grade students. The differences are small, however, and it is not necessary for our purpose that the sample be exactly equivalent to this particular population. Furthermore, the slight differences in average scores on an intelligence test would probably not significantly affect the distribution of scores on the clerical test which is not measuring intelligence to any great extent.

As a further description of our sample, Table 3 shows the percentage of each age represented in each of the five grade groups.⁴ The age is that at the nearest birthday.

A still further description of our sample was obtained by determining from the pupil's statements the occupation of the father and then distributing these occupations according to the categories of the Occupational Rating Scale⁵ of the University

⁴The reader may be interested in noting the resemblances between this distribution and that which Terman and Merrill used for the standardization of the revised Stanford-Binet test. L. M. Terman and M. A. Merrill, *Measuring Intelligence* (New York: Houghton-Mifflin, 1937), p. 17.

⁵Florence L. Goodenough and John E. Anderson, *Experimental Child Psychology* (New York: Appleton-Century, 1931), pp. 501-12.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 3

AGE GRADE DISTRIBUTION IN PERCENTAGES FOR CASES INCLUDED
IN THIS STUDY

Grade	N	Age									
		12	13	14	15	16	17	18	19	20	21
VIII	572	6	35	43	10	4	1				
IX	659		3	39	42	12	4				
X	788			5	35	41	16	2	1		
XI	808				7	36	40	12	4	1	
XII	1,077					5	38	38	14	3	1

of Minnesota Institute of Child Welfare. A total of 3,347 of our 3,904 cases were so classified from occupations as given by the pupils whose fathers were living, employed in an urban community, and not on relief. There were 557 cases not classified, and these included some for which the information was inadequate. The elimination of these groups from the classification, therefore, tends to give a slightly distorted picture of our sample, weighting it for the upper socio-economic levels. Such elimination was necessary for comparative purposes with figures available for the United States population and a similar urban community, Minneapolis.

Table 4 presents the results of this classification for each grade and for those of our total sample who were classified. Comparisons may be made first with the distribution for Minneapolis. Our sample appears to be slightly skewed towards the higher occupational levels. Part of this is accounted for in the number who were unclassified. Despite that, 46 per cent of our sample have fathers in the upper three occupational groups, and 49 per cent of the Minneapolis male population are in these three groups. There are more striking discrepancies, however, when our sample is compared with that for the male population of the United States as a whole. It is

TABLE 4
PARENTAL OCCUPATIONS OF CASES IN THIS STUDY DISTRIBUTION OF THE KNOWN OCCUPATIONS OF FATHERS WHO WERE LIVING, EMPLOYED
IN AN URBAN COMMUNITY, AND NOT ON RELIEF

Occupational Class	Grade VIII		Grade IX		Grade X		Grade XI		Grade XII		Total Sample		Male Pop	
	N	%	N	%	N	%	N	%	N	%	N	%	Mpls*	U S
I Professional	40	8.2	27	4.7	27	4.1	39	5.6	78	8.4	211	6.3	54	3.7
II Semi-professional and managerial	100	20.6	84	14.8	93	14.0	86	12.3	152	16.5	515	15.4	63	6.1
III Clerical, skilled trades, retail business	118	24.2	126	22.1	175	26.4	171	24.5	230	24.7	820	24.5	373	17.7
V Semi-skilled, minor clerical and business	176	36.1	256	45.0	274	41.5	309	44.3	366	39.4	1,581	41.3	243	36.2
VI Slightly skilled	37	7.6	45	7.9	62	9.4	58	8.5	68	7.3	270	8.0	149	13.3
VII Day laborers, urban and rural	16	3.3	31	5.5	32	4.8	35	5.0	36	3.9	150	4.5	118	23.0
Total classified	487	100.0	569	100.0	663	100.0	698	100.0	950	100.0	5,547	100.0	100.0	100.0

*Figures calculated from the 1930 census. From F. L. Goodenough, *The Kuhlman-Binet Tests for Children of Pre-School Age* (Minneapolis: University of Minnesota Press, 1928), Table 1, p. 17.

unlikely that norms developed in a single community would be typical of all communities. Norms developed in a single locality, when well described, are more useful than those derived from many diverse populations, a combination of which may not be typical of any one situation.

Table 5 presents the condensed grade norms⁶ for the decile points for boys and girls separately in grades eight through twelve on the number checking (Test I) and name checking (Test II) tests of the Minnesota Vocational Test for Clerical Workers. These norms were derived from ogive curves constructed from the distributions of cases including all ages within each grade. The number of cases and description of subjects at each grade level have been reported earlier in this article. The test user who wishes to apply this test to subjects at these grade levels should consider whether his population is similar to the one used for calculation of these grade norms.

Some persons will prefer age norms, and for this reason we are presenting in Table 6 age norms for these same subjects who were enrolled in grades eight through twelve. We recommend the use of the grade norms whenever possible, however, as they represent actual grade populations which have been described. The age norms do not include an entirely representative sampling at these ages since we included only those pupils enrolled in school in grades eight through twelve. Furthermore, unequal numbers were selected at the various grade levels. Actually, however, the similarities between age and grade norms are more striking than the differences. The grade eight norms are similar to the age norms for fourteen-year-old pupils, for example. Also notice the striking resemblance between the norms for eleventh grade and seventeen-year-old pupils.

In conclusion, it is suggested that the grade norms should

⁶Complete percentile norms are available in reference 10 and from the test distributors, The Psychological Corporation, 522 Fifth Avenue, New York City. For all practical purposes, however, the less refined interpretations of the test scores will be all that are required.

NORMS FOR MINNESOTA CILRICAI TEST

TABLE 5
CONDENSED GRADE NORMS FOR BOYS AND GIRLS IN GRADES EIGHT THROUGH TWELVE ON
TESTS I AND II OF THE MINNESOTA VOLUNTARY TEST FOR CILRICAI WORKERS

Deciles	Score		Score	
	Test I	Test II	Test I	Test II
Grade VIII				
	Males (N = 284)		Females (N = 288)	
10	110	115	105	100
9	108	99	121	120
8	101	91	114	104
7	93	86	110	102
6	89	81	105	96
5	85	77	100	92
4	80	72	95	87
3	76	67	90	82
2	72	63	85	78
1	65	57	76	70
0	50	15	25	15
Grade IX				
	Males (N = 332)		Females (N = 327)	
10	155	165	180	180
9	120	118	131	128
8	110	105	122	119
7	101	94	116	111
6	98	88	111	105
5	91	81	107	101
4	98	79	101	95
3	81	71	93	91
2	78	69	91	84
1	71	60	83	75
0	10	30	50	10
Grade X				
	Males (N = 372)		Females (N = 416)	
10	165	170	185	180
9	123	121	141	140
8	117	109	133	130
7	106	102	127	121
6	102	96	120	113
5	98	91	114	106
4	92	86	109	100
3	86	81	101	95
2	81	73	97	89
1	75	65	87	79
0	55	35	55	35
Grade XI				
	Males (N = 381)		Females (N = 427)	
10	180	195	190	190
9	141	132	149	147
8	121	118	110	137
7	111	110	133	129
6	106	103	128	124
5	101	97	122	118
4	96	91	117	113
3	91	86	111	106
2	85	79	106	100
1	76	71	97	90
0	45	40	55	55
Grade XII				
	Males (N = 539)		Females (N = 538)	
10	185	180	195	195
9	137	111	151	153
8	128	127	142	145
7	118	119	135	137
6	111	111	129	131
5	106	101	122	124
4	100	98	118	117
3	95	92	112	109
2	89	85	106	102
1	82	78	97	92
0	50	40	30	10

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 6
CONDENSED AGE NORMS FOR BOYS AND GIRLS IN GRADES EIGHT THROUGH TWELVE ON
TESTS I AND II OF THE MINNESOTA VOCATIONAL TEST FOR CUPRICAL WORKERS

Deciles	Score		Score	
	Test I	Test II	Test I	Test II
Age 14				
	Males (N = 246)		Females (N = 297)	
10	142	162	162	172
9	114	115	131	129
8	107	103	122	118
7	101	94	117	111
6	95	89	112	105
5	89	83	107	100
4	85	78	103	95
3	80	73	97	90
2	75	68	92	83
1	69	62	84	75
0	42	32	27	37
Age 15				
	Males (N = 323)		Females (N = 345)	
10	162	167	182	192
9	124	121	138	139
8	110	111	130	129
7	105	101	121	120
6	99	94	115	112
5	94	89	110	106
4	90	83	105	100
3	86	78	100	95
2	80	70	94	88
1	72	60	85	77
0	52	37	52	47
Age 16				
	Males (N = 362)		Females (N = 411)	
10	182	192	187	187
9	127	135	145	146
8	117	115	137	136
7	109	105	131	128
6	104	97	123	121
5	100	91	118	113
4	94	86	112	105
3	88	81	107	100
2	83	75	100	93
1	77	68	90	84
0	47	37	57	47

NORMS FOR MINNESOTA CLERICAL TEST

TABLE 6 (Cont.)

Deciles	Score		Score	
	Test I	Test II	Test I	Test II
Age 17				
	Males (N = 433)		Females (N = 454)	
10	177	177	192	182
9	135	137	150	150
8	125	122	140	140
7	117	112	133	132
6	110	105	128	125
5	104	100	122	118
4	99	94	116	111
3	93	88	110	104
2	86	81	105	97
1	78	71	96	83
0	47	37	57	47
Age 18				
	Males (N = 268)		Females (N = 259)	
10	182	177	177	172
9	135	130	152	150
8	124	122	142	139
7	114	111	134	131
6	107	105	128	125
5	102	99	122	118
4	98	90	118	111
3	94	85	113	105
2	88	79	107	97
1	79	70	97	87
0	52	42	32	62

be useful in junior and senior high schools and commercial business colleges which are concerned with the distribution of their pupils to the commercial classes upon the basis of aptitude for clerical work rather than upon the basis of such factors as lack of aptitude for college training. This test with its norms should contribute toward a more scientific and wiser counseling of pupils based upon all of the pertinent and available information. Norms for adults employed in clerical occupations also might be employed in judging a pupil's clerical aptitude.

REFERENCES

- 1 Andrew, Dorothy M "An Analysis of the Minnesota Vocational Test for Clerical Workers I," *Journal of Applied Psychology*, XXI (1937), 18-47
- 2 Andrew, Dorothy M "An Analysis of the Minnesota Vocational Test for Clerical Workers II," *Journal of Applied Psychology*, XXI (1937), 139-72
- 3 Andrew, Dorothy M "An Analysis of the Minnesota Vocational Test for Clerical Workers," Ph D thesis, University of Minnesota Library, 1935
- 4 Andrew, Dorothy M "The Construction and Standardization of a Test for File Clerks," Master's thesis, University of Minnesota Library, 1931
- 5 Andrew, Dorothy M and Paterson, Donald G "Measured Characteristics of Clerical Workers," *Bulletin of the Employment Stabilization Research Institute*, University of Minnesota, III, 1 (1934), 60
- 6 Andrew, Dorothy M. and Paterson, Donald G "Minnesota Vocational Test for Clerical Workers Manual of Directions" New York The Psychological Corporation, 522 Fifth Avenue, 1939
- 7 Candee, Beatrice and Blum, Milton "A New Scoring System for the Minnesota Clerical Test," *Psychological Bulletin*, XXXIV (1937), 545
- 8 Dvorak, Beatrice "Differential Occupational Ability Patterns," *Bulletin of the Employment Stabilization Research Institute*, University of Minnesota, III 8 (1935), 46
- 9 Green, Helen J, Berman, I R, Paterson, D G, and Triabue, M R "A Manual of Selected Occupational Tests for Use in Public Employment Offices," *Bulletin of the Employment Stabilization Research Institute*, University of Minnesota, II 3 (1933), 31.
- 10 Schneidler, Gwendolen G "Further Studies in Clerical Aptitude," Ph D thesis, University of Minnesota Library, 1940
- 11 Stead, William H, Shattle, Carroll L, and Associates *Occupational Counseling Techniques* New York American Book Co., 1940 273 pages

EXAMINING EXAMINERS

NORMAN J. POWELL

New York City Civil Service Commission

THE EXAMINATION of applicants and the establishment of lists of persons eligible for appointment to professional positions in any school system is a most important task. In New York City, the examining work is performed by a board of seven examiners selected as the result of competitive examination given by the Municipal Civil Service Commission. In view of the considerable current interest in the matter of examinations for teacher and administrative personnel, a somewhat detailed description of the procedures used by the New York City Civil Service Commission in the most recent test given for examiners may be of suggestive value.

It should be noted that civil service examinations, by their nature, are subject to peculiar and serious difficulties. Since examinations cannot be repeated it is not ordinarily practicable to obtain evidence as to the validity of specific test material; the selection and use of such material must rest largely upon judgment and indirect evidence. The passing mark is often arbitrarily set by law — usually at the 70 or 75 per cent point — necessitating nice judgment on the part of examiners as to the difficulty of the test material in relation to the calibre of the applicants, if the examiners do not judge the situation accurately, it is then necessary to resort to the transformation of scores.

Since the examinations are given as a public service and the system depends upon public approval, the examinations must give the appearance of being just and reasonable, even

to the person who knows nothing about examinations. Finally, elements of the examination procedure are subject to appeal and review by the courts, it must, therefore, be defensible before judges who know nothing about examination techniques. No model procedure has yet been developed to meet all needs and situations. The following account presents one careful and painstaking approach to the specific problem at hand.

Adopted in 1937, there is a statutory requirement in the New York Education Law to the effect that applicants for examiner positions must be college or university graduates and possess at least five years of public school teaching experience. To this minimum qualification, the Civil Service Commission added the further requirement of three years of administrative experience in the field of education. Applicants were also required to be not more than 49 years of age at the time of filing application. In consequence of a state law, only residents of New York State were permitted to compete in the examination.

A total of 114 applications was received. Of these, 88 were adjudged as meeting the education, experience, age, and residence requirements.

The Written Test

Consisting of Dean Ned H. Dearborn of New York University, President Paul Klapper of Queens College and Director Paul M. Mort of the Advanced School of Education, Teachers College, Columbia University, a special committee was designated by the New York City Civil Service Commission to prepare the written test.¹

The written test, weighted 4, together with an oral test, weighted 2, and an evaluation of candidates' training and experience, weighted 4, comprised the entire examination. In order to be allowed to take the oral test, candidates had to pass the written and, in addition, had to pass the oral test to be eligible to have their training and experience evaluated. Only 61 of the 88 qualified persons appeared for the written

¹The writer served as aide to each of the committees who worked with the examiner test.

EXAMINING EXAMINERS

test Three applicants withdrew after part of the examination, leaving a total of 58 candidates

Divided into four equally weighted parts and with a mark of 65 per cent in each part as well as a general written average of 75 per cent required to pass, the written test was in neither traditional objective nor essay form The abilities to be measured did not appear to lend themselves to usual objective test treatment The precise abilities taken for measurement may be exemplified by reference to both the questions used in the test and the directions given to candidates

In Part I, for which the candidate was allowed three hours, the applicant was informed

"In rating this paper consideration will be given to clarity in defining the problem, cogency of facts used, orderly presentation of thought, conciseness of expression, and the general effectiveness of the analysis and discussion "

A single three-hour essay was to be written on one of five problems of which the following is illustrative

"It has been suggested that an examining board concerned with improvement of its techniques should maintain a research division

"Analyze this proposal discussing the functions, the organization, the personnel, the values, and the limitations of such a division "

Another example is

"It is maintained that in examinations for promotion there must be full recognition of the contributions which the candidate made in the subordinate position

"Discuss this policy of recognition from the point of view of an examiner in a school system "

The second part, requiring four hours, consisted of 25 technical questions The directions were similar to those given for the first part An illustrative question is

"A reliability coefficient of .55 can be said to be typical for rating personality traits by ordinary judgment methods Is this coefficient high or low? What basis do you have for your answer? What difficulties are involved in the interpretation?"

In another type of question requiring a longer response, the candidate was directed to assume the establishment of a new supervisory position in the Department of Education, was given the duties of the position, the requirements for

which had been set up, and was asked to state whether he agreed with the requirements established, whether there should be additional requirements, and to give a critical analysis of the statements regarding the oral test to be given.

The form of Parts III and IV was similar. In each the applicant was told:

"This part of the examination is a test of your ability to analyze a given problem, to document your position, and to employ sound reasoning. The examiners will be concerned with your ability to present evidence, not with the nature of your attitudes. You are required to demonstrate the depth and breadth of your scholarship in answering these questions."

Four hours were allowed for the completion of each part. There were 40 items in each. Typical questions in Part III are:

"5 'Education is a phase of civilization, not the whole.' What are the implications of this statement for education as one of the societal agencies?"

"9 'No philosophy of education is fundamental until it is based on sociology—not on physiology, not even on psychology, but on sociology.' Is this a valid statement? Why?"

"15 Is it possible for education to be non-partisan? Why?"

Examples of the questions in the fourth part are:

"5 'The very fact of contact between two cultures tends to engender features new to both.' What basis is there for this statement?"

"17 'Inertia conditions the solid framework of society and makes culture possible.' Is this a valid statement? Why?"

"32 'As a group, the aged are increasing faster than the general population.' Enumerate three highly significant social effects."

An effort was made to eliminate the deficiencies of customary essay testing and to introduce the major advantages of the objective test, while retaining the virtues of essay testing. It may be pointed out that the considerable length of the examination made possible both intensive and extensive sampling. The type of item used in Parts III and IV has been subjected to quantitative analysis by the Social Science Research Council with the finding that it is an "extraordinarily useful" instrument. Dr. Brigham is of the opinion that

EXAMINING EXAMINERS

the kind of examination question utilized in the third and fourth parts of the test measures "breadth of background," though the directions in the test under consideration here state that depth of scholarship is also to be demonstrated.² Certainly Part I approaches closely the appraisal of depth aspects and Part II probes depth to a somewhat lesser degree than it evaluates breadth.

The central difficulty involved in essay examinations is unreliability of rating. To promote objectivity in rating the last two parts of the test, the extent of the response by candidates was constricted temporally and therefore spatially. A seven-point rating scale was employed in which the characteristics of best, mediocre, and poor answers were recorded to serve as guides for the awarding of credits. Definite key answers were formulated to all the questions in Parts I and II of the test and a scale for the allocation of credits was set up. In all parts of the test, rating keys were constructed by two or three examiners in conference, partially by reference to relevant literature or other sources, partially by examining candidates' answers to provide a realistic scoring basis. Marking was performed independently by two or three examiners who, after the completion of the scoring, compared their ratings. All discrepancies except those of a trivial character were noted and candidates' answers were reread to find an equitable base for agreement by the raters as to the mark merited by the specific answer. Final ratings were the mean of the individual examiners' marks.

For the type of examination employed in Parts III and IV, a rating reliability of .87 for total score has been found for a 50-question, four-hour social science test.³ In a 40-question examination of similar form given for promotion to Captain, Department of Correction, New York City, the correlation for total test score between two raters was .93. Split-half reliability adjusted by the Spearman-Brown formula was

²C. C. Brigham, *Examining Fellowship Applicants* (Princeton: Princeton University Press, 1935), pp. 22-3.

³*Ibid.*, p. 14.

92, while the standard error of measurement was 3.30.⁴ There appears to be fairly substantial evidence that the kind of examination constructed is satisfactorily reliable.

Much of the basis upon which the widespread belief in essay test unreliability rests seems to be a derivative of experimental findings arising from biased investigations. The bias is a result of rating without the use of keys so that differences between raters are differences between judgments as to the nature of correct responses and the magnitude of the credits to be awarded to partially correct answers as well as divergencies in the appraisal of particular responses. It appears exceedingly probable that there would be differences of opinion among experts in the rating even of many multiple-choice questions if the experts were not provided with a scoring key. In the present instance the formulation of key answers and rating scales, frequent conferences among raters, and the use of several raters tend greatly to eliminate unreliability of rating in each part of the written test.

The essential, significant characteristic of a test is its validity. Reliability is only of incidental importance since a test may be reliable without being valid but cannot be valid unless it is also reliable. Unfortunately, in the written test as in the oral and experience measures, it is not possible to compute a validity coefficient in terms of an acceptable criterion of ability on the job. No satisfactory criterion exists, only one candidate was appointed subsequent to the examination.

A validity judgment may, however, be predicated upon two elements, the backgrounds of the special examining committee and the reasonableness of the appearance of the examination. Both factors support the belief that the written test is valid. The examining panel consisted of prominent educators highly experienced in the selection of personnel. Also, the written test ranged widely over many subjects apparently pertinent to examining work, and its length was sufficiently great to make validity an exceedingly probable attribute of the test.

⁴Bureau of Research, New York City Civil Service Commission, "Selection of Captains in the New York City Department of Correction," *Public Personnel Quarterly*, I, 1, 6-7.

EXAMINING EXAMINERS

The final factor of great importance is the differentiating capacity of the test. In terms of maxima of 100 per cent, the applicants' scores are set forth below:

	Part I	Part II	Part III	Part IV
Mean	62.1	53.7	47.1	50.0
Standard Deviation	15.1	11.6	12.8	10.0
Highest Score	94.6	71.4	68.6	77.1
Lowest Score	35.0	26.3	21.4	25.7
Range	59.6	45.1	47.2	51.4

Test scores separate well among candidates. Applicants are distributed over approximately 50 percentage points, about half the total possible range. It is the middle half of the scale which is occupied by candidates' scores. The range is roughly from 25 to 75 per cent except for Part I where scores are distinctly higher. The highest mark for the written test combining all four parts was 76.7 per cent. The passing mark was 75 per cent.

It follows that either the written test was too difficult or the candidates were too poorly equipped. Either conclusion suggests the desirability of transmuting original scores into higher marks. If the test was too difficult, some of the failures should be passing persons. If the applicants are defective, the condition is unfortunate but largely the product of the rigid statutory requirements limiting applicants to particular groups. An extensive publicity campaign had insured that all or practically all qualified persons were aware of the opportunity to compete in the examination.

The adjustment of marks involves the necessity for determining the nature of the transmutation process. In each test part, the mean was taken as the point of reference and denominated 75 per cent. Distances in standard deviation units above and below the mean fixed the precise percentages awarded to candidates.

Thus, of the 58 candidates, 29 were passed in the written test. The purpose of the examination was to place on an eligible list those persons who appeared to be qualified for the position of examiner. With this guiding principle in mind, it was considered desirable that the better half of the candi-

dates taking the written test be given the opportunity to submit to further examination. It was considered that all those in the upper half of the group had demonstrated the possession of a comparatively acceptable minimum of scholarship. Rescaling, then, involved taking one point of reference instead of another. It was believed that incompetents who managed to slip by in this process would be caught in the tests which were to follow. An oral test then was administered to the 29 persons who had survived the written.

The Oral Test

The test was given in two parts equally weighted and separated in time by about six weeks. The first part was designed to measure technical competence, the second set out to appraise judgment, clearness and quickness of comprehension, manner, appearance, and speech.

In Part I of the technical-oral test, the 29 persons who had passed the written examination were divided into six groups, five with five persons and the sixth with four persons. For each group a demonstration oral examination was enacted. The demonstration orals were ostensibly given for a particular job. The particular positions for which the demonstrations were held were teacher of English in the high schools, teacher of economics in the high schools, psychologist, research assistant, elementary school principal, and director of adult education.

One group of candidates observed one demonstration, a second group observed another, and so on. In each case both demonstration examiner and subject were members of the examining division of the Commission. Into each demonstration certain defects and virtues were introduced both with regard to the demonstration examiner and the demonstration subject. Demonstrations were written and planned in advance. Candidates were required to rate both participants in the interview which was enacted. Each demonstration lasted for about one-half hour. Candidates were permitted to take notes while the demonstration was in progress and then allowed an additional fifteen minutes for note taking.

EXAMINING EXAMINERS

Following the demonstration, candidates retired to an adjacent room and were summoned individually for an oral examination before the examining panel. This oral examination lasted for not less than one-half hour.

Candidates were rated by the panel in accordance with a set of directions which had been formulated in advance and in accordance with criteria prepared prior to the demonstration and adjusted after the demonstration to fit the performance which had been observed. The members of the panel viewed the demonstration at the same time as the candidates in order to be able to adjust accurately the criteria for rating to accord with the demonstration observed by the candidates.

The ratings received by candidates were determined by the ratings they had given to participants in the demonstration and by the adequacy of the support they were able to adduce for their ratings. The candidate's evaluation of the demonstration examiner was required to be supported by observations on the examiner's attitudes toward the subject, his skill in questioning, and the general conduct of the interview. The evaluation of the demonstration subject was required to be supported by observations on speech, manner, judgment, and appearance.

Of the 29 who had taken Part I of the oral, 16 qualified to proceed to the second part. That an applicant was "qualified to proceed" did not necessarily mean that he passed Part I, since a general average of 70 per cent in the oral test as a whole was required in order to pass. For example, candidates who obtained 60 per cent in the first part of the oral proceeded to the second part of the oral with the possibility of passing the entire oral test only if they obtained a score of 80 per cent on the second part, which would give them the required average of 70 per cent. The marks received by the 16 candidates who qualified in the first part of the oral were

<i>Mark</i>	<i>Frequency</i>
60.0	3
60.8	2
62.1	1

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

62.5	3
63.3	1
65.0	1
65.8	1
66.7	1
74.2	1
75.8	2

Only three persons obtained scores of 70 per cent or better in the first part of the oral. It is indicated, then, that the large majority of the candidates who appeared for the second part of the oral had already exhibited mediocrity with regard to technical competence.

There is a sizable discrepancy in score between Part I and Part II in only two cases. In one case, a candidate who had obtained 75.8 in the first part received 60.8 in the second. In the other case, a candidate who had obtained 74.2 in Part I received 59.0 in Part II. The point of these data is that candidates were consistently poor. The fact of the matter is that most of the candidates performed in mediocre fashion in the first part of the oral and merely confirmed their mediocrity in the second part of the oral, already having shown inferiority in the written test.

The 16 individuals who took Part II of the technical-oral test were divided into eight groups of two. For each group of two persons, examined separately in a particular half day, two types of situations were set up. In the first type of situation, the candidate was directed to assume that he had been appointed to the position of examiner and that he had been serving in this position for about five years. The candidate was told that he would be visited by a person with whom he was to talk for about half an hour and that he was to conduct this interview as naturally and effectively as he could. For this examination, one person, Professor Robert K. Speer of New York University, acted as the visitor and assumed a different role for each group of two candidates. The roles assumed were representative of a parent association, assistant examiner, colleague on the board, reporter, representative of the

EXAMINING EXAMINERS

Civil Service Commission, visitor from Sioux City interested in teaching personnel, failed candidate, and representative of a teacher training institution

A second type of situation was established after the conclusion of the candidate's interview with the visitor. This consisted of three or more questions. Examples of the first kind of question are

Can education be reconstructed through research?

How would the adoption of a particular philosophy of examining in New York City affect educational practice and thinking throughout the whole country?

What do you consider to be the major virtues (or defects) of our educational system in the United States?

For the second question, the candidate was given a quotation, asked to tell whether he agreed or disagreed with the quotation and what the implications were of his agreement or disagreement for the work of the Board of Examiners in New York City. Some typical quotations are:

"The danger which comes from emphasizing the significance of contemporary changes is that hasty and unsound revisions will be made in the curriculum."

"In the solution to educational problems lies the solution to all social problems."

"The main function of education is to perpetuate democracy."

"Adult education should be limited to educable adults."

For the third question, the candidate was required to talk for several minutes on any topic which he deemed to have implications for the work of the Board of Examiners. This was followed, where appropriate, by having members of the panel question the candidate directly in order to obtain clarification or amplification of one or more points made by the candidate. It is noted that members of the panel were free to ask questions of the candidate at any time in order to explore a statement by the candidate.

The direct questioning by the panel was introduced by having the candidate talk for several minutes on his experience and background mainly in order to give the candidate an opportunity to "warm up" prior to the questioning. The ratings given to candidates were made in accordance with writ-

ten directions adopted by the examining panel. In the situation in which the candidate assumed that he was already an examiner, the following rating criteria were employed: soundness of position taken, cogency of discussion, clarity of discussion, penetration of treatment, time taken for effective organization of responses to the visitor, manner and attitude adopted toward the visitor, quality of speech, and appearance. In the situation involving more direct questioning, the following criteria were used: importance of material selected, soundness of position taken, relevance of material selected, clarity of presentation, penetration of treatment, quality of speech, manner and attitude adopted toward the panel, time taken for effective organization of responses, and appearance.

Information for the rating of the five factors was supplied both by the direct questioning and the assumed examiner situations. In the rating of the five factors, a rating scale was employed which ranged from 0 to 100 per cent. The standards set were

"Unacceptable candidates should be given ratings below 60 per cent. Ratings between 60 per cent and 75 per cent should be given to candidates whose characteristics considered in this part of the test are only very slightly inferior in level to those of a high-grade examiner. Ratings above 75 per cent should be given to candidates who undoubtedly possess a high level of the characteristics defined in this part of the examination."

There were three examiners in the first part of the technical-oral. Joseph G. Cohen, Director of the Division of Graduate Studies, Brooklyn College, Ned H. Dearborn, Dean of the Division of General Education, New York University, Margaret V. Kiely, Dean, Queens College.

Because the second part was less susceptible to objective rating, the examining committee was increased to five in order to minimize subjectivity. Ned H. Dearborn of New York University; Willard S. Elsbree, whose special field is teacher personnel, of Teachers College, Columbia University; Margaret V. Kiely of Queens College; Jesse H. Newlon of Teachers College, Columbia University; Ordway Tead, President, Board of Higher Education in New York City.

EXAMINING EXAMINERS

The traditional deficiencies of oral tests are well known and include in civil service examining the difficulty of achieving both the fact and the appearance of satisfactory reliability and validity. Appearances are of considerable importance in public personnel administration. Not only must the examination be an effective instrument but also it must avoid the impression of being arbitrary, unfair, or capricious even though it is none of these in fact. The difficulty is generally that of describing adequately the basis for ratings and of connecting clearly the rating scale with the candidate's performance in the determination of marks. It must be proved that marks are accurate and unbiased. This necessity was recognized and met by setting forth in writing the nature of the scoring scales, criteria, and standards used, and by keeping stenotype and phonographic records of all questions and answers. The effort to have the examining panels be rather large and representative of diverse educational viewpoints and to have them consist of leaders in the profession was also considered to contribute toward the objective of coupling seeming with actual validity.

The technical requisites for reliability seem to be present. The average intercorrelation of the examiners' ratings on the first part was .91, making an estimated reliability for their composite ratings of .97, by the Spearman-Brown formula. The average intercorrelation among the examiners on the second part was .785, making the estimated reliability for their composite ratings .95. The average difference among examiners in Part I of the oral is 2.2, in Part II the average difference is 5.4. Scores were in five-point units as 50, 55, 60, 65, 70, so that the disparity in grades is about half of one point on the Part I rating scale and about one point on the scale in Part II.

Since quantitative appraisal of the validity of the oral tests is impossible, the problem must again be approached logically. Validity refers to the degree to which a test measures what it sets out to measure. The oral attempted to evaluate ability to judge applicants for educational positions,

to analyze weaknesses and strengths in oral examining methods, to deal with visitors, to display good judgment and comprehension, and to exhibit a satisfactory appearance, manner, and speech. Situations were formulated with the explicit purpose of measuring these factors, all of which appear to be significant samples of the examining task, so that from the viewpoint of job analysis the oral appears to be acceptably valid.

The Combined Scores

When the scores for both parts of the oral were combined, it was found that only one candidate had achieved a passing rating. At this point in the examination, however, adjustment of marks to pass a greater number was deemed undesirable. There are several reasons for not transmuting marks in the oral test. In the oral, the identity of candidates is known. In the written, identity is concealed by having candidates enter their application numbers in place of their names.⁶ To transmute marks where identity of applicants is known is to make possible the charge of manipulation. Further, the written was followed by other tests able to weed out the unfit, the oral was to be followed by an experience test in which every person admitted to the examination was certain to receive a passing mark because all possessed the prescribed minimum education and experience qualifications. Moreover, very substantial opportunity had been afforded to aspirants for the position of examiner to prove themselves. Applicants had been examined at four separate occasions in the written test and at two different times in the oral. Finally, the matter of standards is highly relevant in deciding whether or not to rescale marks.

The position of examiner is of the utmost importance in a school system. The examiner is responsible for the selection of educational personnel and therefore, in a large measure, for the quality of the teaching done and the manner in

⁶The practice of the New York City Commission is to affix rating numbers to all written test papers and to detach the application number from answer sheets. The applicant knows his application number, but not his rating number.

EXAMINING EXAMINERS

which the youth of the city is taught and molded. The position pays \$11,000 a year and is held for life after a six-month probationary period which is not made effective since no appointee to this position has ever been discharged after probationary appointment. It also must be borne in mind that educational practice in New York City affects to a degree educational practice in the remainder of the country. It seems reasonable to believe that under these conditions a high standard is desirable for this position. The position was taken by the Commission that passing only one of 58 persons taking an examination of this type is not evidence of an unjustifiably high set of standards when the number of jobs to be filled is very small.

A great row arose after the eligible list of a single name was published. It would be interesting and instructive to take up the controversy in detail, but such a discussion belongs elsewhere. Some of the objections can be laid to a lack of understanding of fundamental measurement principles.

The examination was reviewed three times by the courts and once by a committee on manifest errors established by the New York City Civil Service Commission. First of the many and varied interpretations of the data came with the appointment of the committee on manifest errors to hear and judge candidates' appeals. The usual procedure of the Commission is to refer all appeals to a board of three members of the Commission staff. In view of the importance of this examination, however, and the necessity of eliminating any suspicion of bias, the special panel was constituted. Its personnel consisted of Arthur A. Ballantine, noted lawyer and Undersecretary of the United States Treasury under President Herbert Hoover, Charles J. Pieper, Professor of Science Education and head of the Department of Science Education at New York University; William F. Russell, Dean of Teachers College, Columbia University. In a report dated May 3, 1938, the Commission found "no manifest error in the examining methodology or in the constitution of the examining panel," stated that "the pass mark was set neither too high nor too

low in relation to the level of competency required," and concluded that there was "no manifest error in the rating of any candidate."

Suit to invalidate the test was then brought by seven of the failed candidates. It was held by the New York Supreme Court "that the technical-oral test against which the principal assault was made was meticulously prepared and impartially administered, that every safeguard to insure fairness and equality of competition was provided, and that the standards used in rating the competitors were in legal contemplation objective and reviewable."

The failed candidates had greater success with the State Appellate Division. Five justices of the Appellate Division concurred in finding the oral examination invalid. The justices disagreed quite strongly in regard to the selection of the ground upon which to rest their conclusion. One stated that it was illegal to limit the eligible list to one name; a second was impressed with the "comparative incompetence" of the sole passing applicant, others interpreted the evidence to point to the intrusion of ideological considerations in the technical-oral test.

The final word came with the decision of the Court of Appeals which disagreed with the Appellate Division as to why the oral test was illegal but agreed that the test should be given all over again.

Following these vicissitudes, the New York City Civil Service Commission held a new oral test in 1940. This time, three candidates were passed. The applicant who had been the only one to qualify in the previous test was included among the three who were successful in the new one.

NEW CRITERIA FOR OLD

F. R. SARBIN AND L. S. BORDIN¹

University of Minnesota

IF ALL the literature on the prediction of college grades were to be assembled in one place, the outstanding characteristic would be the almost universal agreement that correlation coefficients higher than .70 are practically impossible with existing methods. As a matter of fact, Segel has collected over a hundred such studies only to discover that the median predictive validities of high school scholarship, tests of general achievement or aptitude, and tests of specific aptitudes or achievements were .54, .44, and .37 respectively.²

In studying the factors which are responsible for these relatively low coefficients, our attention is immediately focused on the nature of the criterion—the honor-point ratio. Commonly used by colleges and universities as an index of the student's achievement, this summary figure represents attainment in many different kinds of courses taught by various kinds of teachers with different standards of measurement.

Two characteristics of this criterion are of importance for predictive efficiency, namely, its unreliability and its heterogeneity. The first characteristic, unreliability, has not really been measured effectively, but can be estimated by logical analysis. It is agreed that even with improved methods of measuring attainment in college courses a semi-intuitive, hit-

¹We are indebted to Professor E. G. Williamson for stimulation and advice in the formulation of this paper. We gratefully acknowledge his permission to use part of the data contained in his study *Prediction of Success in the Arts College* to be published in bulletin form by the University of Minnesota.

²David Segel, *Prediction of Success in College* (U. S. Office of Education, Bulletin 1934, No. 15), p. 70. See also Daniel Harris, "Factors Affecting College Grades: A Review of the Literature, 1930-37," *Psychological Bulletin* XXXVII (1940), 125-66.

or-miss judgmental factor still remains in the grading process.² That this would create a measure of unreliability in the individual course grade is undeniable. As long as each teacher has a set of standards, individually derived and reflecting a somewhat unique set of objectives, so long will grades retain their unreliability. When we compound the unreliabilities of the individual course grades—which we do in computing honor-point ratios—it is improbable that the reliability of the final criterion will approach the reliability of the predictors.

If it were possible to establish perfect reliability of course grades in individual subjects and of the honor-point ratio, the second characteristic of the criterion, heterogeneity, would still remain to interfere with prediction. For students who are taking courses in natural sciences, mathematics, social sciences, and languages in varying combinations, the criterion represents a complex of many factors each of which logically ought to be sampled by the components of the predictive battery. It is self-evident that the more complex and heterogeneous the factors in the criterion, the more difficult becomes the task of assembling a predictive test battery which will adequately sample this aggregate without simultaneously introducing into the predictive index other extraneous factors. Our task would be solved if we could assemble a series of pure measures for each component in the criterion. Pure tests, however, have not yet been created. The early promise of the factor analysts that a pure test was possible has not yet been realized.

A word of caution is in order for those who would hasten, after having discussed the unreliability and heterogeneity of the prevailing criterion, to do something about it. The unreliability or reliability of a criterion is only one factor in prediction. Of equal importance are the reliability and validity of the predictive battery. As already indicated, techniques

²This is not to imply that judgments are to be abandoned. By learning to avoid the pitfalls and fallacies in human judgments, teachers can improve the quality and consistency of their ratings. Several writers have treated at some length the common errors in making judgments. See H. E. Buritt, *Principles of Employment Psychology* (Boston: Houghton-Mifflin, 1926), Chapter II, and M. S. Viteles, *Industrial Psychology* (New York: W. W. Norton Company, 1932), Chapters IX, X.

have not yet been developed for creating tests which will be pure measures of any single factor. Thurstone's utilization of factor methods in his Primary Abilities tests has not yet passed beyond the experimental stage. In fact, first reports have been conflicting.⁴ While the reliabilities of the predictive tests such as the American Council and the Ohio Psychological Examination distribute around .90, these alone do not offer hope for a great deal of improvement in validity. Thus, research ingenuity must be applied to the predictor variables as well as to the criterion.

One final limiting aspect of prediction must be taken into account by the research worker before pitching his aspirations too high. This is the indeterminancy principle that Heisenberg has formulated for prediction in the physical sciences. Present-day thinkers recognize that spontaneous and uncontrolled factors are always present. These cannot be foreseen, they will introduce a measure of error in any forecast. Among such factors to be found in the prediction of academic achievement are momentary motivations such as health conditions, social distractions, sexual distractions, home conflicts, temporary moods, sets, fatigue, and so on. Because of these not readily controllable elements, it would be safe to guess that even with perfectly reliable criteria and with statistically infallible predictive tests, the upper limit of multiple correlation would still not exceed .95. But such a pessimistic outlook need not be discouraging to further research. Much room remains for improvement. The increase in predictive efficiency of a correlation of .70 to one of .95 represents a range of about 41 per cent improvement over non-test estimates.

The crux of the problem of selection and admission of students hinges upon accurate prediction instruments. Prediction serves the purpose of assisting college authorities to

⁴J. M. Stalnaker, "Primary Mental Abilities," *School and Society*, I (1939), 868-72. See also R. G. Bernreuter, "Primary Ability Tests Applied to Engineering Freshmen," *Psychological Bulletin*, XXXVI (1939), 548-49, and William M. Shanner and G. Frederic Kuder, "A Comparative Study of Freshman Week Tests Given to the University of Chicago," *Educational and Psychological Measurement*, I (1941), 85-92.

select students who have a reasonable chance of profiting from the college's offerings. Since the efforts of the test-makers, educational psychologists, and other research workers reached a ceiling at forecasting efficiency of approximately 28 per cent better than non-test prediction, further research may take any of three courses

- 1 further improvement in the reliability and validity of the predictive battery,
- 2 improvement in the reliability of the criterion measures,
- 3 design of a new criterion which will be more predictable and at the same time acceptable to school administrators

At the present time, the first approach appears to be the one least likely to bring success, yet it is the one most frequently selected. With the development of the method of factor analysis hopes were raised for a significant increase in the efficiency of tests. The belief prevailed that with the isolation of factors in a test battery the foundation might be laid for the construction of pure tests which in turn would lead to more accurate prediction. Thus far this promise has remained unfulfilled.⁵ Until now the more significant contribution in test construction has come from the method of inbreeding of test items as utilized by Toops in the construction of the Ohio Psychological Examination.⁶ By means of this continuous process of selection of the most valid and most stable items, the predictive validity of the Ohio test has at times surpassed .60. The promise for further developments from this source is at present greater than from the method of factor analysis. The stimulus for further advances by the method of inbreeding probably will come from studying the contributions of the alternatives in a multiple choice item.⁷ But even with this contribution the prospects for the near future are not

⁵Stalnaker, *op cit*. See also Bernreuter, *op cit*.

⁶H. A. Toops, "The Evolution of the Ohio State University Psychological Test," *Ohio College Association Bulletin* No. 113, March 20, 1939, pp. 2267-311.

⁷G. F. Kuder, *The Construction of Valid Test Items* (Unpublished Dissertation, Ohio State University, June, 1937).

very bright for a large increase in reliability and validity of tests

The second course, increasing the reliability of the criterion measures, sporadically has been the topic of intense discussion in educational circles. As far back as 1913 Starch appealed for more stable grading standards. The major factors which he cited as the cause for instability of marks still are applicable today:

"(1) Differences among standards of different schools, (2) differences among standards of different teachers, (3) differences in the relative values placed by different teachers upon various elements in a paper, and (4) differences due to pure inability to distinguish between closely allied degrees of merit."⁸

In the last decade a new type of emphasis in the grading process has arisen largely through the influence of Tyler⁹ and the Progressive Education Association evaluation work. The efforts of this group have been directed mainly toward the clarification of teachers' aims and objectives and the operational definition of these aims and objectives in terms of observable behavior. These developments have been directed chiefly at the secondary school level in connection with the Eight-Year Study.

The wider application of these principles at the college level may offer some hope for the improvement of prediction. It is assumed that this type of study will lead to a more conscious and a more stable evaluative process which in turn should serve to make grades more reliable. Some believe that greater homogeneity in the objectives of grades also would result from these developments. That is to say, many objectives probably could be identical for different courses. If a core of common objectives could be isolated and evaluated in the same manner in a whole series of courses, then the difficulty of constructing a more efficient test battery would be reduced considerably.

⁸Daniel Starch, "Reliability and Distribution of Grades," *Science*, XXXVIII (1913), 630.

⁹Ralph W. Tyler, "Needed Research in the Field of Tests and Examinations," *Educational Research Bulletin*, XV (1936), 151-58.

The third approach is most likely, we believe, to bring about significant increases in the predictive efficiency of test batteries and is one that probably would encounter the most opposition from administrators and faculty. If we assume that the present grade criterion of college success lacks adequate predictability and that this deficiency warrants the substitution of a more predictable criterion, then is it not logical to seek such a criterion? The answer can be only in the affirmative. At least a portion of our efforts must be directed at the possibilities of developing a more predictable measure of academic achievement which at the same time will satisfy other needs of the educational program.

But we also must consider the difficulties of such an undertaking and be prepared to surmount them. Over and above the educational and statistical problems are the sociological problems which arise from the nature of our educational society. This society has developed a rigid and inflexible attitude toward marks which is likely to resist any but very strong pressures.

To dislodge the tradition of marks, two forces must be overcome: first, the faculty, who feel that they have a vested interest in assigning grades, and second, parents, who, "indifferent at times to most phases of education, seldom neglect the report card."¹⁰ This rigid adherence to marks has another deleterious effect upon the educational process. Instead of directing their efforts toward mastery of content, many students prepare for grades. Originally designed to serve merely as a record that a student had taken a particular course and had acquired a certain degree of proficiency, grades too often have become the only goal for many students. Foerster has described the situation in pungent terms:

"Once a credit was earned, it was as safe as anything in the world. It would be deposited and indelibly recorded in the registrar's savings bank, while the substance of the course would be, if one wishes, happily forgotten. Each course culminated in a final examination, if one knew one's stuff then, one need never know it again. In a subject like required

¹⁰R. O. Billett, *Provisions for Individual Differences, Marking and Promotion* (U. S. Office of Education, Bulletin 1932, No. 17), p. 459.

NEW CRITERIA FOR OLD

English, a student deficient in ability might, with effort, get a passing grade, and then, without effort, pass into semi-illiteracy, yet the record would show, to the day of doom, that he could read and write passably"¹¹

All this means that institutions of higher learning may have to abandon or modify the traditional marking system and "produce a new convention better than the old."¹² It is thus seen that continued enslavement to traditional marking systems not only interferes with the construction of more effective selection instruments, but also vitiates some of the fundamental objectives of higher education. Therefore, upon the shoulders of the educational administrator falls the responsibility of re-examining the purposes of a marking system—a system that he either implicitly or explicitly has set up as proper. In this re-examination he will be obliged to leave the way open for the substitution of another marking system which will provide optimal satisfaction of these purposes.

Our problem has come into sharp focus: a new standard for gauging achievement in college must be sought. This standard must palpably be superior to teachers' marks and must rest on certain logical and statistical pillars. Our previous discussion of the limitations upon predictive accuracy for college selection purposes has already indicated some of the desirable features: first, the measure should have reliability, second, it should be as homogeneous as possible both with respect to scale and to the nature of the factors included, third, it must have relevancy for the educational objectives to be measured.

Since the final test of the predictability of a criterion will be empirical, we turn to the data that we have to present. At this juncture the nature of the statistical evidence we have obtained forces us to particularize in terms of the liberal arts college, and more specifically, the junior division. The underlying principles, however, can readily be adapted to other college units. Before proceeding with the analysis of the com-

¹¹Noiman Foerster, *The American State University* (Chapel Hill, N. C., University of North Carolina Press, 1937), p. 97.

¹²*Ibid.*, p. 146.

parative predictability of the two types of criteria, a word must be said about the objective of the junior division of the liberal arts college. At the risk of seeming impertinently presumptuous in stating this objective in a word, the authors suggest that the primary purpose of the first two pre-specialization years in the arts college is to provide students with opportunities for cultural growth. Generally speaking, today's liberal arts colleges by and large direct their efforts—sometimes futilely—toward a cultural goal. Although in this context the word culture is to be looked upon with the gravest suspicion, the trend in today's core curricula seems to be away from the hot-house variety of culture for the elite and in the direction of the by-products of the best in science and society for all. In most cases, the American liberal arts college is bent upon providing students with a broad understanding of culture in all of its ramifications.

To be cultured, a man must be "more than an ape-like creature posing under the mask of hastily acquired drawing room manners"¹³. The student must acquire during his pre-specialization years the individual qualities and competences which go into rich and satisfying living, and which give meaning to his experiences as a member of society. This does not imply that a cultural pattern rigidly common to all is the goal of liberal education. To diagoon widely different students into a legion of regimented automatons, each responding in the same way to the same situations, is obviously to be deplored in democratic institutions. As Eckert has phrased it: "Not like minded, but 'free' individuals become the goal of teaching"¹⁴. Idiosyncratic behavior remains as an outstanding desideratum of liberal education.

At this point one of two approaches is immediately apparent for evaluating these cultural objectives. We may retreat to the traditional methods of evaluation—teachers' grades based on some esoteric combination of improvised testing,

¹³C. J. Warden, *The Emergence of Human Culture* (New York: Macmillan Company, 1936), p. 8.

¹⁴Ruth E. Eckert, "Who Are the Cultured in Our Colleges?" *Educational Record*, January, 1930, pp. 133-35.

dazzling intuitions, and the persistence of the student in attending classes, or we may turn to procedures such as those embodied in certain uniform testing programs

Such a measure of culture could be assembled with the Cooperative General Culture test as a nucleus.¹⁶ Since 1932, the content of the Sophomore Culture test has been considerably expanded, and today the student runs a gamut of tests from mathematics to aesthetic appreciation before entering his junior year. Admittedly, the paper-and-pencil instrument does not sample the whole range of culture, neither does it directly tap the important areas of motivation, attitudes, and values. It is the only method yet devised, however, which has the fundamental characteristics without which scientific measurement in education becomes a farce, a tragedy, or both. If college administrators and faculties are to decide whether the Sophomore Culture test will satisfy their needs, they must weigh it upon scales which carry empirical as well as logical weights. If it is agreed that the predictability of a proposed criterion is one characteristic pertinent to its adoption, the predictability of the Sophomore Culture test becomes a matter of moment.

One might point out that it has a high reliability—coefficients in the .90's are reported in the literature—or that it is constructed so as to give comparable scores, but the final proof of its superior predictability must rest upon obtained correlations with predictive tests. The remainder of this paper is devoted to an exploratory investigation of the predictability of the two criteria we have been discussing—teachers' marks¹⁷ and the Sophomore Culture battery.

The usual technique was employed in assembling a battery

¹⁶The Cooperative General Culture test may be procured from the Cooperative Test Service, 15 Amsterdam Avenue, New York. The other Cooperative tests used in this study may be obtained from the same source.

¹⁷Teachers' marks were transmuted to two year honor-point ratios as follows: for each credit hour in which an A was recorded, three honor points were assigned, for each credit hour of B, two honor points, for each credit hour of C, one honor point, for each credit hour of D, no honor points, and for each credit hour of F (failing) one honor point was subtracted. The honor-point ratio was computed by dividing the total number of honor points earned by the total credit hours earned. The Sophomore Culture test was made up of the following tests in the Cooperative series for 1936: General Culture, English, General Science, Literary Acquaintance.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

of tests for the selection or rejection of applicants. These tests were correlated first with teachers' marks and then with scores on the Sophomore Culture tests. The battery of entrance tests which has the highest correlation with the criterion could thus be used to select future candidates for admission. At the time of the students' entrance into the arts college, scores on the following measures were obtained.

High school percentile rank

Minnesota College Aptitude test (form AM)

Minnesota College Aptitude test (form 1926)

Cooperative English test (Part I, form 1934)

Cooperative Vocabulary test (Part II of English test above)

Cooperative Contemporary Affairs test (form 1934)

The group included in this study was composed of students who entered the arts college of the University of Minnesota as freshmen in the fall of 1934, and who took the Sophomore Culture test in the spring of 1936 in applying for admission into the upper division. The group was composed of 138 students, 56 men and 82 women. Only students were included for whom the 1934 entrance test scores were available and for whom high school percentiles were recorded. The group studied, though not closely representative of entering freshmen, probably was representative of sophomores applying for entrance to the senior division of the arts college. Any limitation in representativeness, however, invalidates no comparisons between different measures within this group.

TABLE 1

CORRELATIONS BETWEEN TWO YEAR HONOR-POINT RATIOS AND INDIVIDUAL MEASURES IN THE FRESHMAN TESTING BATTERY

	Total	Men	Women
High school percentile rank	.52	.57	.55
Contemporary Affairs test	.50	.53	.44
Minnesota College Aptitude test (1926)	.50	.56	.43
Cooperative English test	.41	.50	.43
Minnesota College Aptitude test (AM)	.49	.49	.39
Cooperative Vocabulary test	.35	.40	.30

Table 1 reveals the usual order of correlations between teachers' marks and predictive tests. The best single predictor

NEW CRITERIA FOR OGD

of grades is the high school percentile rank, demonstrating that—to a certain extent—high school teachers and college teachers are influenced by the same factors in assigning grades. The highest correlation in the table is .57, between grades for men and high school percentile ranks. The lowest coefficient, .30, is between college grades for women and the Vocabulary test. The other coefficients fall between these two values.

TABLE 2
CORRELATIONS BETWEEN SOPHOMORE CULTURE TEST AND INDIVIDUAL MEASURES IN
THE FRESHMAN TESTING BATTERY

	Total	Men	Women
Contemporary Affairs test	.81	.81	.82
Minnesota College Aptitude test (1926)	.77	.77	.76
Cooperative Vocabulary test	.68	.63	.72
Minnesota College Aptitude test (AM)	.67	.68	.66
Cooperative English test	.58	.62	.66
High school percentile rank	.29	.43	.21

Contrast these correlation coefficients with those in Table 2. With the single exception of the high school percentile rank, correlations between the Sophomore Culture test and the various measures range from .82 to .58. The Contemporary Affairs test has high predictive value, as have the College Aptitude tests and the Vocabulary test. The nature of the distribution of the English test scores accounts for the lower coefficient for the total group than for either the men or the women. It is especially noteworthy that high school percentile ranks have little predictive value for such a criterion. In terms of forecasting efficiencies for the total group, the highest coefficient in Table 1 corresponds to 15 per cent, while the highest in Table 2 corresponds to 41 per cent.¹⁷

The same trend appears when the multiple correlation coefficients of selected batteries are compared. Table 3 reveals the order of correlation between the two criteria and two sets of entrance tests. Battery A, composed of three measures (high school percentile rank, Minnesota College Aptitude test

¹⁷Forecasting efficiency computed by formula $E = 100 (1 - \sqrt{1 - r^2})$ which gives a measure of the per cent of improvement over non-test prediction. See J. P. Gullford, *Psychometric Methods* (New York: McGraw-Hill, 1936), p. 363.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

—form 1926, and English test), correlates .64 with grades, but .77 with the Sophomore Culture test. The corresponding indexes of forecasting efficiencies are 23 per cent and 36 per cent. When the Contemporary Affairs test is added to the three other measures (Battery B), the correlation with honor-point ratio becomes .67, and with the Sophomore Culture test .86. The corresponding forecasting efficiencies are 26 per cent and 50 per cent better than non-test prediction.

TABLE 3

MULTIPLE CORRELATION COEFFICIENTS BETWEEN BATTERIES OF SELECTED ENTRANCE TESTS AND TWO YEAR HONOR-POINT RATIO, AND SOPHOMORE CULTURE TEST

	Two Year Honor-Point Ratio			Sophomore Culture Test		
	Total	Men	Women	Total	Men	Women
Entrance Battery A*	.64	.67	.64	.77	.73	.77
Entrance Battery B†	.67	.63	.66	.86	.86	.87

*Entrance Battery A—high school percentile rank, Minnesota College Aptitude test (1926), and Cooperative English test.

†Entrance Battery B—high school percentile rank, Minnesota College Aptitude test (1926), Cooperative English test, and Contemporary Affairs test.

The differences between these multiple correlations for the two kinds of criteria were tested for significance. The differences for Battery A were in the area of doubtful validity ($P < .05$ but $> .02$), those for Battery B were well beyond the boundary for significance ($P < .01$).¹⁸

The significant results of this exploratory study in the prediction of college success may be summarized as follows: (a) The substitution of the Sophomore Culture test for the conventional grading system as a criterion of college achievement markedly increases the predictive validities of the standardized entrance tests and markedly decreases the predictive validity of high school grades. (b) The lowest validity coefficients were obtained when high school percentile ranks were correlated with the Sophomore Culture battery. The highest zero order coefficients were obtained in correlating the Sophomore Culture battery with the Contemporary Affairs test. Eckert

¹⁸R. A. Fisher, *Statistical Methods for Research Workers* (7th ed., London: Oliver and Boyd, 1939), p. 209.

reports similar findings. She concludes: "Students most conversant with the achievements and thoughts of the past, and most outstanding in the realm of book-learning, tend on the whole to be those most alert to the contemporary scene"¹⁹

(c) A combination of four entrance measures returned validity coefficients with the Sophomore Culture test corresponding to 50 per cent forecasting efficiency. The Contemporary Affairs test alone correlated higher with the Culture test than did a combination of three measures.

Interpreting these results, the Sophomore Culture test correlates high with the other objective tests because of its close similarity in objectivity of form, its greater relevancy and comprehensiveness, and in the overlapping of the content and ability measures, and correlates low with high-school grades because the latter appraise other areas besides those involved in the test sampling of achievement. Grades in college, conversely, correlate higher with grades in high school, and lower with the standardized tests because they measure areas outside of tested achievement but similar to those measured by high-school grades. This interpretation can be further supported and extended since the correlation between the Sophomore Culture test and the two year honor-point ratio was only moderately high: .58 for men and women combined, .64 for men, and .51 for women. Scholastic grades and the Culture test, even when they presumably sample the same areas of knowledge, certainly do not measure all of the same areas or abilities involved in academic achievement in college.

These results are not without precedent. For example, Frasier and Heilman reported correlations between the Thorndike Intelligence Examination and grades assigned subjectively and objectively. The average coefficients were .45 and .60 respectively.²⁰ For grades in French as assigned in the usual manner, Tharp found a correlation of .47 with the Iowa Place-

¹⁹Eckert, *op cit.*, p. 135

²⁰G. W. Frasier and J. D. Heilman, "Experiments in Teacher College Administration, III. Intelligence Tests," *Educational Administration and Supervision*, XIV (1928), 268-78.

ment test for foreign language aptitude. When an achievement test was used, the correlation jumped to .64.²¹

From our examination of the problem of prediction, we draw the conclusion that a fruitful point of attack is through the substitution of a more reliable and therefore more predictable measure of achievement. This paper has presented data which definitely demonstrates that a pencil-and-paper evaluation instrument such as the Sophomore Culture test is more predictable than the time-honored grade criterion. But it would be foolhardy indeed for the authors to take the next step, that of advocating that this attribute alone justified its substitution for honor-point ratio. This decision lies within the province of the educational administrator. He must decide whether more accurate prediction — a *sine qua non* of all efficient admissions policies — plus the Culture test's degree of relevance is sufficient to outweigh those desirable qualities which may still be claimed for the traditional marking system. In short, he must decide whether this new criterion is more acceptable than the old.

A final word for research. The Sophomore Culture test, in common with other achievement tests, largely measures recall of information.²² That information is only one phase of education must be recognized. Other components of cultural growth — attitudes, values, motivations, goals, and affective experience — must be measured by other instruments. It is hoped that in the not-too-distant future these important outcomes of education can be appraised with sufficient accuracy so that we may know how well the American college functions as the vehicle of culture.

²¹J. B. Tharp, "Sectioning Classes in Romance Languages," *Modern Language Journal*, XII (1927), 95-114.

²²B. E. Cureton, "Evaluation or Guidance—A Report of the 1939 Sophomore Testing Program," *Journal of Experimental Education*, VIII (1940), 308-40.

A FACTOR ANALYSIS OF A NON-VERBAL REASONING TEST

ROBERT I. BLAKELY

Social Security Board

SOME time ago Dr Andrew W. Brown and the author constructed a "Non-Verbal Reasoning Test" for use at the high school level. A preliminary report of its construction is being published by *The Journal of Educational Psychology*. The present article concerns itself with the results of a factor analysis of the intercorrelations between the subtests rather than with the actual standardization of the test.

The test was constructed with the idea that it should measure in a non-verbal manner the higher intellectual processes of comprehension, mental alertness, deductive reasoning, inductive reasoning, and spatial relations or analysis. The primary purpose of this study is to isolate and identify any common factors present and to compare them with the expected factors.

Other problems which may be considered in the light of the factor analysis are (a) a comparison of the factorial composition of tests which are variations of Thurstone's tests with the factorial composition, as determined by Thurstone, of the tests he used, (b) a reconsideration of the perennial problem of the existence of a general factor of mental ability; (c) the comparison of the factors found in this group of tests with factors found in analyses of other tests, (d) a further examination of various methods of ascertaining the number of factors which should be taken out of a correlation matrix.

All tests are time-limit tests and were introduced by four exercises which were explained by the examiner. They were presented in the order listed.

1 *Manikin*—a page of pied figures of little men. The figures are simple line drawings with variations in the positions of arms and legs. The problem is to draw a ring around each manikin which is exactly like a model at the top of the page. It was thought that this test might be saturated with the Perceptual Speed factor. The Spearman-Brown corrected reliability is .81.

2 *Identical Patterns*—12 rows of patterns formed by overlapping geometrical forms. The first pattern of each row is separated from the others by a heavy vertical line. The patterns are in 12 variations each composed of two circles and two right triangles. The same size forms are used in each variation, the differences being due to relative positions of the components and whether the forms are solid or dotted lines. Each row contains one or more patterns exactly like the first one in the row, and the problem is to place a mark under each pattern which is exactly like the first one in its respective row. It was thought that this test would be a variation of Thurstone's *Identical Forms* test and consequently loaded with the Perceptual Speed factor. The Spearman-Brown corrected reliability is .98.

3 *Fitting Parts*—each item consists of a solid black geometrical form, which has been cut into three parts, and four outlined figures, one of which is the same size and shape as the black figure which was cut. The problem is to indicate that one of the outline forms into which the solid black pieces could be made to fit exactly. Discrimination of both size and shape is involved for each item. It was thought that possibly the factor Visualization or Space was involved in the solution of this test. The Spearman-Brown corrected reliability of the 12-item test is .47.

4 *Opposite Sides*—each item consists of three drawings identical in size and shape. The problem is to select the drawing in each item which is a mirror image of the other

two drawings. Each drawing is a little pennant the shape of a non-isosceles right triangle and may be rotated in any position. It was thought that possibly Space and Induction might be used in the solution of this test. There is no really parallel form to this test although the idea was adopted from Thurstone's *Flags* test. The Spearman-Brown corrected reliability is .88.

5. *Code*—a code consisting of eight boxes divided in half is placed at the top of the test. Each box has a unique group of squares and circles in the top half and an unusual group of triangles in the bottom half. Below the "code" are five rows of the little boxes, some exactly like the boxes in the code and some with incorrect pairing of the symbols. The problem is to place a line under each box which is different from the code. It was thought that the test might contain the Perceptual factor. The Spearman-Brown corrected reliability is .96.

6. *Circle Grouping*—each item consists of four boxes containing little groups of circles. The grouping varies from box to box. One circle in each of the first three boxes is blackened according to a system. The problem is to discover that system and apply it in blackening a circle in the fourth box. It was thought that possibly Induction would be involved in solving this test. The Spearman-Brown corrected reliability for the 12-item test is .98.

7. *Form Series*—this test is the usual series type with only three meaningless forms used in combination. One figure in each row is omitted and a blank inserted. The problem is to indicate which form belongs in the blank. It was thought that Deductive Reasoning or Inductive Reasoning would be involved in the solution of this test. The corrected Spearman-Brown reliability of the 22-item test is .86.

8. *Circle Reasoning*—a variation of the *Marks* test used by Thurstone as a measure of Inductive Reasoning. There are five rows of groups of circles and dashes. The grouping changes from row to row. One circle in each of the first

four rows is blackened according to a rule. The problem is to find the rule and apply it in blackening a circle in the fifth row. It was assumed that this test would contain Induction. The corrected reliability is .94.

9. *Form Relations*—this test is a parallel form of Thurstone's *Pattern Analogies* test. The problem is to indicate one of five choices which bears the same relation to the third figure as the second bears to the first. Inductive Reasoning or Deductive Reasoning was assumed to be necessary for the solution of this test. The corrected reliability is .97.

10. *Form Reasoning*—at the top of the test is a table showing how any two of seven forms could be combined to equal another one of the seven. Each item consists of three of the forms in a row. The task is to combine the first two forms according to the table and then combine the resulting form with the third to equal another form, the final result to be indicated by underlining one of five choices. It was thought that possibly Deductive Reasoning would be used to solve these problems. The Spearman-Brown corrected reliability for the 12-item test is .98.

The Subjects

The subjects were 286 high school pupils from a school in a suburb of Chicago. All tests were given by two experienced examiners in a well-lighted room. All tests were administered in one 40-minute period. Eighty per cent of the whole group was between 15 and 18 years of age. The mean Otis I.Q. was 114. About 54 per cent of the group were boys. No sex difference was found for combined scores on the whole test. No grade difference was statistically significant. The correlation of total test score with chronological age was $-.13$ for the age range of this group.

The Factor Analysis

The table of intercorrelations (Table 1) was computed with the aid of *Computing Diagrams for the Tetrachoric Correlation Coefficient* (2). Correlations obtained in this

manner are considered by Thurstone (6, p 58) to be applicable to factor analysis. In effect the scores are normalized in the process of correlation.

The factors (Table 2) were extracted by the Thurstone centroid methods. Here the problem of the number of factors

TABLE 1
INTRACORRELATIONS OF TESTS

Variable	1	2	3	4	5	6	7	8	9	10
Manikin		24	27	24	38	19	13	19	22	19
Identical Patterns	24		08	17	22	46	16	15	33	24
Fitting Parts	27	08		17	22	20	13	10	20	22
Opposite Sides	24	17	17		15	25	38	32	39	31
Code	38	22	22	15		26	22	25	35	38
Circle Grouping	19	46	20	25	26		48	50	53	49
Form Series	13	16	13	38	22	48		35	52	54
Circle Reasoning	19	15	10	32	25	50	35		55	38
Form Relations	22	33	20	39	35	53	52	55		40
Form Reasoning	19	24	22	31	38	49	54	38	40	

TABLE 2
CENTROID MATRIX (F)

Variable	Code No	Factors				
		I	II	III	IV	V
Manikin	1	438	—435	—183	—083	—069
Identical Patterns	2	452	—141	274	263	—200
Fitting Parts	3	335	—212	—101	—140	087
Opposite Sides	4	499	100	—163	—112	—242
Code	5	506	—297	—055	—079	130
Circle Grouping	6	701	138	296	272	109
Form Series	7	622	377	110	—274	—117
Circle Reasoning	8	602	281	—252	238	205
Form Relations	9	728	181	—154	177	—093
Form Reasoning	10	665	119	166	—251	239

appeared. Two methods of determining the number of factors had been tried by the author (1) previously with some degree of success. One of these, Tucker's empirical criterion, gave negative results in the present case. The other, Coombs' criterion (3) postulates that in a 10-variable problem, the last factor of value will leave a table of residuals which, when signs are changed, will contain more than 31 negative entries with a standard error of five. Table 3 shows the application of Coombs' criterion to this analysis.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 3
COOMBS' CRITERION

Factor	Negatives
1	24
2	33
3	24
4	28
5	35

It was obvious from the number of relatively large residuals remaining in the table after the second factor was extracted that there were more than two factors in the table. This was borne out in the subsequent analysis, which was carried to five factors. The indication that the fifth factor was the last one of value seems to have been verified in the analysis. The standard deviation of the fifth factor residuals before sign change is .028, which is considerably smaller than the standard error of a zero correlation for a population of 286.

For the rotation of factors in order to secure bounding hyperplanes, Thurstone's method of lengthened vectors was used (4). The criteria of maximizing the number of zeros and rotating to a postulated positive manifold were the determinants for direction of rotation. Seven rotations were necessary and a "clean-up" rotation with actual length vectors was made. The rotated factorial matrix is given in Table 4. The rotational matrix of direction cosines is given in Table 5. The intercorrelations between the rotated factors are presented in Table 6.

TABLE 4
ROTATED FACTORIAL MATRIX (1A)

Variable	Code No	Factor				
		A	B	C	D	E
Manikin	1	.582	.075	.004	.192	.054
Identical Patterns	2	.092	.547	-.016	.239	.014
Fitting Parts	3	.345	-.041	-.009	.265	.005
Opposite Sides	4	.132	.028	.160	.313	.408
Code	5	.440	.067	.004	.394	-.062
Circle Grouping	6	-.076	.436	.161	.641	-.073
Form Series	7	-.141	.040	.016	.639	.453
Circle Reasoning	8	-.010	-.046	.561	.518	.026
Form Relations	9	.076	.162	.415	.507	.244
Form Reasoning	10	.080	.021	-.053	.766	.071

ANALYSIS OF NON-VERBAL REASONING TEST

TABLE 5
TRANSFORMATION MATRIX (A)

Centroid Axis	Reference Vector				
	A	B	C	D	E
I	287	247	207	801	203
II	-859	-243	361	248	379
III	-380	671	-690	234	-224
IV	-184	521	582	-257	-436
V	033	-399	111	421	-759

TABLE 6
CORRELATIONS BETWEEN NORMALS TO THE PLANES (A' A)

Plane	Plane				
	A	B	C	D	E
A	1 000				
B	-.084	1 001			
C	-.092	-.241	1 000		
D	-.011	-.007	-.009	1 001	
E	-.127	-.117	-.005	-.003	1 001

Even a cursory glance at the rotated matrix will show that the factorial composition of the tests is not so simple as had been hoped for.

Factor "A" has three variables with significant projections and all the others are essentially zero. These are

1. <i>Manikin</i>	58
3. <i>Fitting Parts</i>	35
5. <i>Code</i>	44

Either one of two interpretations could be placed on this factor. It might be considered to be Space as has been described by Thurstone (6), the author (1), and others. Under this interpretation it would seem that the grasping of spatial relations of the arms and legs of the manikins was of more importance than the quick perception of small differences. It would appear also, that the quick comparison of the code with the stimuli in the *Code* test was not so important in solving the problem as the grasping of the relationship between the two halves of the individual elements.

The other interpretation which could be placed on this factor is that it is Perceptual Speed, or rather mental alert-

ness, as distinguished from Perceptual Discrimination. Under this interpretation the ability would involve the quick change of response from item to item with only the simplest discrimination necessary. Thurstone's factor "9" in his study of Hyde Park High School in Chicago seems to have some of the characteristics of factor "A" (5). In this case, the test *Scattered "x"s* had the highest loading. The *Manikin* test has the simplest discrimination level and the *Fitting Parts* test the most complex of those listed. The author prefers this latter interpretation.

Factor "B" has two tests which have significant loadings:

2	<i>Identical Patterns</i>	.	55
6	<i>Circle Grouping</i>	.	44

It seems obvious that this factor corresponds to Thurstone's (6) Perceptual Speed factor, but we shall call it Perceptual Discrimination to distinguish it from factor "A". The difference here is that the emphasis is on analytic perception in which a fine discrimination must be made rather than on speedy response to a simple stimulus. Speed is of importance, but in the subjects used the differences in the mental process of perceptual discrimination will contribute more to performance variance than will simple speed.

At first glance it seems surprising that *Circle Grouping* is high on this factor. However, a careful subjective analysis of the test will indicate that the problems involved are more those of perceptual discrimination than of induction. The figures are complex but the rules to be brought out are simple. For example, one of the items has the middle dot blackened in a group of three, which is apparent even at a glance, so that the problem resolves into finding the correct group in the response square. This takes a discriminatory ability evidently slightly below that required for *Identical Patterns*.

Factor "C" has two variables with significant projections:

8	<i>Circle Reasoning</i>	.	56
9	<i>Form Relations</i>	42

ANALYSIS OF NON-VERBAL REASONING TEST

Both of these tests are variations of tests used by Thurstone in his studies of the primary mental abilities and have been interpreted to contain Induction, or Inductive Reasoning. This interpretation is suitable in the present case. The apparent paradox that test 8 contains Induction while test 6, in which a supposedly similar function is involved, does not may be resolved when an inspection is made of the tests themselves. The primary problem in test 8 is to find a rule by which the problem may be solved while in test 6 the main problem, as has been said before, is to find the response group rather than the rule.

Factor "D" is an orthogonal factor which was set up by making its normal perpendicular to the normals of all the other planes. This was necessary as one dimension of the five-dimensional system could not be identified by a bounding hyperplane because of lack of variables with zero projections in that dimension. It is the same type of problem as was encountered by the author in a former study (1).

All the variables have projections on this factor which are probably significant. The relative amount of projection seems to increase with the greater complexity of the mental function involved. The tenth test, *Form Reasoning*, which involves the synthesis of geometrical figures according to established rules (not unlike arithmetic), has much the highest saturation of the factor.

The obvious comment, and one that must be reckoned with, is that this factor represents "general intelligence," or Spearman's factor "g." As has been said before, there is nothing in the Thurstone method of analysis which denies that such a general factor exists or implies that it would not show up if present. However, in regard to the nature of the present factor, there can be little doubt that it is "general" for this battery of tests and is not an effect of maturation or lack of differentiation of ability due to the youth of the subjects. What it is called—*comprehension, understanding, mental efficiency, or intelligence*—is beside the point. Due to the popular misconceptions and scientific vagueness of the last

term, it probably would be better to adopt some other name

It should be understood that the author is of the opinion that the above-mentioned effect of an augmented general factor due to lack of maturation is applicable to situations in which the subjects are immature, but that such a factor does not account for appreciable distortion in the present case. It is not denied that such a general factor is present in tests given to children, but it seems probable that the general factor, if it exists in such a case, is unduly emphasized by the maturation curves of the abilities

Another interpretation which might be placed on factor "D" is that it is Deductive Reasoning, which in each test requires that the subject must base his conclusions or responses on certain facts which are presented in the test item. However, this is probably another aspect of the foregoing discussion

Factor "E" has significant loadings for two tests and a possibly significant loading for a third

4	<i>Opposite Sides</i>	.	41
7	<i>Form Series</i>	. . .	45
9	<i>Form Relations</i>		24

This factor apparently corresponds with none of the factors previously identified by Thurstone and his associates. However, it may possibly represent Deductive Reasoning as "series" tests have been found by Thurstone (5) to contain a component of Deductive Reasoning. The same is true of the form relations type of test. The relationship of the *Opposite Sides* test to such an interpretation is not immediately apparent. Assuming that one might consider two figures in each item of the *Opposite Sides* test as facts to be compared and from which a conclusion might be drawn concerning the third figure, i.e., whether it is different from the first two or like one of them, then it might be thought to involve Deduction. In the *Form Series* test the symbols presented are facts from which a conclusion must be drawn concerning the missing figure. The conclusion is definitely limited to three alternatives

ANALYSIS OF NON-VERBAL REASONING TEST

each of which might be tried in turn. In the *Form Relations* test the problem might be approached by trying to find the rule involved, which would be Induction, or by substituting the possible answers one at a time and testing the resulting equation. This latter process might be considered to be Deductive Reasoning and insofar as it were used would cause the test to show a loading on the Deduction factor.

No definite conclusion can be made as to the identity of Factor "E," but tentatively it may be called Deductive Reasoning.

Despite the fact that the factorial composition of some of the tests varies somewhat from what was originally supposed, it seems that the tests, as a group, do measure some of the higher mental processes of reasoning. From amount of projection on the general factor, it would seem that the tests saturated with Perceptual Speed are the poorest measures of the higher intellectual processes. It would appear that test number 9, *Form Relations*, which has significant projections on three factors, is probably the best general test of all the reasoning processes. Test 10, *Form Reasoning*, is the best test of the general factor which might be considered to be synonymous with comprehension or mental efficiency or intelligence. The test, *Identical Patterns*, seems to be saturated with the factor Perceptual Discrimination, which is interpreted quite similarly to Thurstone's factor of Perceptual Speed, and is consistent with Thurstone's (5) test of *Identical Forms*, which is parallel in process. The test, *Circle Reasoning*, a variation of Thurstone's (5) *Marks* test, is similar in factorial composition to the latter. The *Form Relations* test seems to have a heterogeneous factorial makeup, as was also found by Thurstone (6).

The factors identified seem to be consistent with those identified by Thurstone (6,5) except for the general factor. It is necessary to investigate these tests in a larger battery before an interpretation can be adequately applied to the general factor. This factor has some characteristics similar to those found by the author (1) in factor "D" in a "Reanalysis

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

of a Test of the Theory of Two Factors" The factor Perceptual Speed also seems similar to the factor "C" in the latter study

The factors have been found to be practically uncorrelated, the highest correlation, that between factors "B" and "C," being only 14 degrees off orthogonality This is probably within chance variation and no significance is attached to it.

REFERENCES

- 1 Blakey, R. I. "A Reanalysis of a Test of the Theory of Two Factors," *Psychometrika*, II (1940), 121-36
2. Chesire, L., Saffin, M., and Thurstone, L. L. *Computing Diagrams for the Tetrachoric Correlation Coefficient*. Chicago University of Chicago Press, 1933. 59 pages
- 3 Coombs, Clyde. Unpublished paper read before the American Psychological Association, September, 1940.
- 4 Thurstone, L. L. "A New Rotational Method in Factor Analysis," *Psychometrika*, III (1938), 199-218
- 5 Thurstone, L. L. "Experimental Study of Simple Structure," *Psychometrika*, II (1940), 153-68
- 6 Thurstone, L. L. *Primary Mental Abilities*. Chicago University of Chicago Press, 1938. 121 pages

NEW TESTS*

California Capacity Questionnaire, by Elizabeth T. Sullivan, Willis W. Clark, and Ernest W. Tiegs. 1941. For high school and college students, and adults. Time, 30 minutes. Forms A and B, 75¢ per 25, 25¢ per specimen set. Published by the California Test Bureau, 3636 Beverly Boulevard, Los Angeles, California.

California Test of Personality, by Louis P. Thorpe, Willis W. Clark, and Ernest W. Tiegs. 1940. One form each for primary, elementary, intermediate, secondary, and adult levels. Time, about 45 minutes for each series. Primary series for grades 1-3, elementary series for grades 4-9, intermediate series for grades 7-10; secondary series for grades 9-14, adult series, \$1.00 per 25 of each series, 25¢ per specimen set of each series. Published by the California Test Bureau, 3636 Beverly Boulevard, Los Angeles, California.

Cooperative Community Affairs Test, by Roy A. Price and Robert F. Steadman. 1941. Time, 30 minutes. Form R, \$3.50 per 100, 25¢ per specimen set. Published by the Cooperative Test Service, 15 Amsterdam Avenue, New York, New York.

Cooperative Literary Comprehension and Appreciation Test, by Hyman Eigerman, Mary Willis, and Frederick B. Davis. 1941. For upper high school and college classes. Time, 40 minutes. Form R, \$4.50 per 100, 25¢ per speci-

*Publishers and authors of new tests are requested to send copies to The Editor, *Educational and Psychological Measurement*, Box 766, Alexandria, Va.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

men set Published by the Cooperative Test Service, 15 Amsterdam Avenue, New York, New York

Cooperative Science Test, by John G. Zimmerman and Richard E. Watson 1941 For grades 7, 8, and 9 Time, 80 minutes Form R; \$5.50 per 100; 25¢ per specimen set. Published by the Cooperative Test Service, 15 Amsterdam Avenue, New York, New York

Cooperative Social Studies Test, by Agatha Townsend and Mary Willis 1941 For grades 7, 8, and 9 Time, 80 minutes Form R, \$5.50 per 100, 25¢ per specimen set. Published by the Cooperative Test Service, 15 Amsterdam Avenue, New York, New York

Dunlap Academic Preference Blank, by Jack W. Dunlap 1940 For grades 7, 8, and 9. Forms A and B, 90¢ per 25, 20¢ per specimen set. Published by the World Book Company, Yonkers, New York

Eames Eye Test, by Thomas H. Eames 1940 \$3.50 for examiner's kit, 65¢ per 25 individual record cards. Published by the World Book Company, Yonkers, New York

Examination for the Measurement of the Efficiency of Mental Functioning, by Harriet Babcock and Lydia Levy 1940 One form, set of test materials, \$11.20, record blanks, \$2.30 per 25, \$6.90 per 100. Published by C. H. Stoelting Company, 424 North Homan Avenue, Chicago, Illinois

Fourth Grade Geography Test, by Zoe A. Thralls, George Miller, and Marguerite Uttley 1940. For use at the end of the fourth grade Time, 35 minutes One form, 8¢ per test, 4¢ per manual, 20¢ per scoring stencil. Published by McKnight and McKnight, Bloomington, Illinois

NEW TESTS

Hills Economics Test, by John R. Hills 1940 For high school and college students Time, 40 minutes One form, 50¢ per 25, 15¢ per specimen set Published by Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas

Kansas Vocabulary Test, by H. E. Schiimmel, O. M. Rasmussen, Anna Huebert, and D. J. Tate 1940 For grades 4 to 8 Forms A and B, 40¢ per 25, 15¢ per specimen set Published by Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas

Kirkpatrick Chemistry Test, by Ernest Kirkpatrick 1940 For high school students Time, 40 minutes One form, 60¢ per 25, 15¢ per specimen set Published by Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas

Kniss World History Test, by F. Roscoe Kniss 1940 For high school students Time, 50 minutes Forms A and B, \$1.30 per 25c, 20¢ per specimen set Published by the World Book Company, Yonkers, New York

Mechanical Comprehension Test, by George K. Bennett 1940 For male high school students and adults Time, about 25 minutes One form, \$2.50 per 25 booklets and answer sheets, 25¢ per specimen set Published by the Psychological Corporation, 522 Fifth Avenue, New York, New York

Minnesota Personality Scale, by John G. Dailey and Walter J. McNamara 1941 For upper high school and college students Time, about 45 minutes Separate question

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

booklets for men and women, answer sheet can be used with either question booklet, scorable by International Test Scoring Machine; \$1.50 per 25 question booklets, 75¢ per 25 answer sheets, 35¢ per specimen set. Published by the Psychological Corporation, 522 Fifth Avenue, New York, New York

Mordy-Schrammel American Government Test, by F. E. Mordy and H. E. Schrammel. 1940. For high school and college students. Time, 40 minutes. One form, 50¢ per 25, 15¢ per specimen set. Published by Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas

Mordy-Schrammel Constitution Test, by F. E. Mordy and H. E. Schrammel. 1940. For high school and college students. Time, 40 minutes. One form, 50¢ per 25, 15¢ per specimen set. Published by the Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas

Peabody Library Information Test, by Louis Shores and Joseph E. Moore. 1940. One form each for college, high school, and elementary school levels. Time, 30 minutes. College level one form, \$1.00 per 25. High school level one form, 75¢ per 25. Elementary school level one form, 60¢ per 25, 20¢ per specimen set. Published by the Educational Test Bureau, 720 Washington Avenue, S.E., Minneapolis, Minnesota.

Rasmussen Trigonometry Test, by O. M. Rasmussen and O. J. Peterson. 1940. For high school and college students. Time, 40 minutes. One form, 50¢ per 25; 15¢ per specimen set. Published by Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas

NEW TESTS

Stanford Achievement Test, by Truman L. Kelley, Lewis M. Teiman, and Giles M. Ruch. 1941. Forms D and E for each of primary, intermediate, and advanced levels from grades 2 to 9. Primary Battery, for grades 2 and 3. Time, 50 minutes, \$1.10 per 25, 20¢ per specimen set. Intermediate Battery—Complete, for grades 4 to 6. Time, 150 minutes, \$2.00 per 25, 40¢ per specimen set. Advanced Battery—Complete, for grades 7 to 9. Time, 150 minutes, \$2.00 per 25, 40¢ per specimen set. Published by the World Book Company, Yonkers, New York.

Tate Economic Geography Test, by D. J. Tate and G. A. Buzzaid. 1940. For high school and college students. Time, 50 minutes. Forms A and B, 50¢ per 25, 15¢ per specimen set. Published by Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas.

Tinsler-Ainett Health Knowledge Test, by V. T. Tinsler, C. E. Ainett, Jr., and H. E. Schrammel. 1940. For grades 9 to 12 and college. Time, 50 minutes. Forms A and B, 50¢ per 25, 15¢ per specimen set. Published by Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas.

Turse Shorthand Aptitude Test, by Paul L. Turse. 1940. For use with high school students before enrolling in shorthand courses. Time, 45 minutes. One form, \$1.30 per 25, 10¢ per specimen set. Published by the World Book Company, Yonkers, New York.

Vocational Inventory, by Curtis G. Gentry. 1940. For high school and college students, and adults. Time, about 150 minutes. One form, 15¢ for vocational inventory, indi-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

vidual analysis report, and individual score tabulation sheet, 25¢ per sample set. Published by the Educational Test Bureau, 720 Washington Avenue, S E , Minneapolis, Minnesota

MEASUREMENT ABSTRACTS*

Adkins, Dorothy C and Kuder, G Frederic "The Relation of Primary Mental Abilities to Activity Preferences" *Psychometrika*, V (1940), 251-62

The relations of abilities, as measured by Thurstone's Tests for Primary Mental Abilities, to activity preferences, as measured by Kuder's Preference Record, are investigated for a population of 512 university freshmen. Ability profiles for contrasted groups on each preference scale reveal relatively slight overlapping between the two sets of measures, although the apparent trends are reasonable. The Pearson intercorrelation coefficients of all pairs of measures involved were determined. Implications of the findings in relation to theory and to educational and vocational guidance are indicated (Courtesy *Psychometrika*)

Allison, G and Barnett, A "Freshman Psychological Examination Scores as Related to Size of High Schools" *Journal of Applied Psychology*, XXIV (1940), 651-52

Quantitative and linguistic scores of 1,083 college freshmen on the 1938 edition of the A.C.E. Test were analyzed with reference to the size of the high schools from which they graduated. For three size-groups, statistically significant differences in means were found in five of six comparisons. Means tend to increase with enrollment but there is much overlapping. *W A Varvel*

Anderson, H A and Traxler, A E "The Reliability of the Reading of an English Essay Test." Part II *School Review*, XLVIII (1940), 521-30

*Edited by Professor Forrest A. Kingsbury

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Factual notes were prepared on two themes "The Discovery of Gold in California" (Form A), and "The Pony Express" (Form B). A group of 281 pupils in the University High School of the University of Chicago were given the two forms at one year's interval with instructions to expand the material into a two-hour essay. The essays were graded on a sixty-point scale with the following weights for the separate factors: completeness (6), spelling (6), punctuation (6), language errors (6), coherence between main divisions (10), organization of paragraphs (10), and organization of essay sentences (10). On rereading 70 essays of each form the grades of a skilled reader showed correlations of 0.93 ± 0.016 and 0.937 ± 0.010 for the two forms, two readers, on first scoring of 25 papers, showed correlations of 0.859 ± 0.035 and 0.898 ± 0.026 for the two forms. For individual factors, no correlation was below .80. Growth in language ability may be indicated by an average gain of 3.3 points from Form A to Form B for 281 pupils. The results are not deemed conclusive but only suggestive of the desirability of experimentation with essay-test procedures. *J. E. Karlin*

Babitz, Milton and Keys, Noel. "A Method for Approximating the Average Inter-Correlation Coefficient by Correlating the Parts with the Sum of the Parts." *Psychometrika* V (1940), 283-88

It is noted that the average inter-item correlation, which represents the internal consistency of a test, yields a unique estimate of test reliability. A close approximation to this average is given by a formula which requires the correlation of each item with the total score and the standard deviation of each item. The formula is especially useful in those instances where the number of items is small and where the variation in item sigmas should not be neglected. (Courtesy *Psychometrika*)

Benton, A. L. and Periy, J. D. "A Study of the Predictive Value of the Stanford Scientific Aptitude Test (Zyve)." *Journal of Psychology*, X (1940), 309-12
Scores on the Stanford Scientific Aptitude Test and the

MEASUREMENT ABSTRACTS

A C E Psychological Examination (1934-35) together with course grades for 43 students over a period of three to four years were used in an investigation of the predictive value of the Aptitude Test. The average score on the A C E was approximately one sigma above the mean for the 1935 freshmen. Correlations of course grades for scientific and non-scientific courses with the Aptitude Test and the A.C.E. Test were about + .35, there being no significant difference between the sets of correlations. The coefficients of correlation of the 11 subtests and average grades in all college courses "were all quite low". The authors suggest "that the test has a certain limited value in prognosticating the scholastic achievement of freshman and sophomore students." *Harold Bechtold*

Buros, Oscar Kissen, Editor. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J. The Mental Measurements Yearbook, pp. 674 + xxxiii. 1941.

The first part of the *Yearbook* contains reviews of new tests as well as of selected older tests. There are 524 tests listed. Most of these are reviewed by two or three reviewers. The second part lists 368 books and pamphlets in the measurement field and excerpts from reviews of them which have been published in various journals.

Cast, B. M. D. "The Efficiency of Different Methods of Marking English Composition." Part II. *British Journal of Educational Psychology*, X (1940), 49-60.

Forty English compositions were marked by 12 examiners by four different methods: (1) the examiner's own habitual method, (2) the method of general impression; (3) Burt's analytic method (allotting separate marks for specified points or qualities), (4) Hartog's achievement method. The P-technique (correlation of persons) was combined with Burt's summation method for a factorial analysis of the correlations between examiners, this resulted in: (a) a general factor (representing the best approximation to the "true marks") accounting for 50 per cent of the variance; (b) a dichotomous factor of examiners marking better by analytic methods or by intu-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

itive or impressionistic methods. The methods of marking for general use are here found to be in order of preference: the "analytic" method, the method of general impression, the examiner's habitual method, and Hartog's achievement method. *J. E. Kailm*

Daniel, C. "Statistically Significant Differences in Observed Per Cents." *Journal of Applied Psychology*, XXIV (1940), 826-30.

A table gives the amount by which a per cent A observed in one sample must exceed a per cent B observed in another sample of the same size to be significant at the 0.05 level. It is presented for different values of B and for samples from 20 to 1,000. Various conditions are stated and the meaning and use of the table discussed. *W. A. Varvel*

Davis, F. B. "The Interpretation of I Q's Derived from the 1937 Revision of the Stanford-Binet Scales." *Journal of Applied Psychology*, XXIV (1940), 595-604.

The author presents a table of equivalent values for I Q's from the 1916 and 1937 revisions of the Stanford-Binet and suggests a new classification of I Q's based on the 1937 form. The method by which the equivalency was calculated is discussed. The suggested classification of I Q's provides a series of equal steps or gradations of brightness. *W. A. Varvel*

Dongan, K. E. and Gory, A. E. "Selecting Unskilled Laborers in Cincinnati." *Public Personnel Review*, I, No. 3 (1940), 43-50.

Job analyses were made of jobs for unskilled laborers as eligible lists became needed. It was agreed that the ability to read and write, a good physique, intelligence, experience, and an age range of 21 to 45 or 50 were required for the jobs.

An examination for waste collector included a practical test calling for repeating a demonstration given by regular workers, an evaluation of training and experience, and an oral interview. A test for street cleaners was composed of 75 multiple-choice items on arithmetic, vocabulary, reasoning, and

MEASUREMENT ABSTRACTS

general information. These questions were put in the language of laborers.

Examining for unskilled labor positions has gone on only since February, 1940. The departments, however, believe they are getting better workers.

Diessel, Paul L. "Some Remarks on the Kuder-Richardson Reliability Coefficient." *Psychometrika*, V (1940), 305-10.

The Kuder-Richardson reliability coefficient is derived in a manner independent of that originally given. Various alternative forms applicable to special situations are exhibited with the purpose of making them available to others interested in using this formula. A simplification in computation is suggested for use with a calculating machine. (Courtesy *Psychometrika*.)

Feigelson, George A. "The Application of Sheppard's Correction for Grouping." *Psychometrika*, VI (1941), 21-7.

This paper attempts to show in a non-mathematical way the influence of grouping on standard deviations and correlations, and advances empirical evidence to illustrate with what accuracy values corrected for grouping by Sheppard's correction approximate values obtained from ungrouped data when the distributions are continuous. This inquiry gained its initial stimulus from the observation that many standard deviations and correlations reported by students of psychology and education are uncorrected for grouping and that frequently errors attributed to the grouping of data are not small when compared with errors of sampling. (Courtesy *Psychometrika*.)

Godard, R. H. and Lindquist, E. F. "An Empirical Study of the Effect of Heterogeneous Within-Groups Variance upon Certain F-Tests of Significance in Analysis of Variance." *Psychometrika*, V (1940), 263-74.

In the application of the analysis of variance to data obtained in educational methods experiments which involve several classes of several schools, one assumption is that of homogeneity in the variances of pupil scores from school to

school. It is shown that such variances on representative educational achievement tests are heterogeneous. The effects of this heterogeneity upon the F-tests of significance commonly employed in methods experiments are investigated by comparing the actual distribution of F values for a large number of "experiments" involving marked heterogeneity with a theoretical distribution based on the assumption of homogeneity. Although the findings, which vary somewhat with the type of variance ratio, are not entirely conclusive, they apparently demonstrate that departure from homogeneity does not invalidate the use of the customary F-tests for evaluating results of the typical methods experiment. (Courtesy *Psychometrika*)

Goodenough, Florence L. and Maurel, Katharine M. "The Relative Potency of the Nursery School and the Statistical Laboratory in Boosting the I Q." *Journal of Educational Psychology*, XXXI (1940), 541-49

This study recomputed data obtained at the Minnesota Nursery School by those statistical procedures generally employed in the Iowa statistical laboratory. In the Iowa procedure, cases were grouped according to initial I Q, instead of paternal occupation. This recomputation of data, which when handled properly showed no effect of nursery school training upon the I Q, gave results similar to those reported from Iowa. A difference in I Q appeared for children who remained at home as well as for nursery school children. The authors conclude that the previously reported differences are the result of fallacious statistical treatment rather than being an educational phenomenon. *D A Peterson*

Gulford, J. P. "The Phi Coefficient and Chi Square as Indices of Item Validity." *Psychometrika*, VI (1941), 11-9

Two new methods of item analysis are described. One involves the computation of the ϕ coefficient (correlation of a fourfold point distribution) and the other involves chi square. The only data required are the proportions of passing individuals in the upper and lower criterion groups, for the

MEASUREMENT ABSTRACTS

determination of ϕ , and in addition, N, for the determination of chi square. Abacs are presented for graphic solution of the two indices of validity, and tests of significance are provided (Courtesy *Psychometrika*)

Jenkins, R. L. "Considerations Relative to the Selection of an Index of Intelligence" *Journal of Educational Psychology*, XXXI (1940), 527-40

The test-retest stability of the I Q and the P C (Hein's personal constant) are compared in terms of Binet test ratings for 1,774 cases. The group was weighted with retarded children. Comparisons of all adjacent tests were made. Regression of both I Q's and P C's toward the mean on retest was found with marked drops in the P C's of very bright children. "The P C. appears to offer no advantage over the I Q for the children of the middle-age group" and appears to be slightly inferior to the I Q at the lower age levels.

The rationale underlying the two statistics are considered. Dispersions in intelligence are assumed in both cases to be proportional to the mental age. The growth function assumed by the I Q and the P C are presented with the point that both curves have one degree of freedom.

It is suggested that a more logical approach would be to express "mental test performance in terms of the sigma value of test score for the chronological age." The assumption is less restrictive than those for the constancy of the I Q or P C; the assumption is "that the relative status of children with respect to intelligence remains constant," which is "implicit in the use of any index of intelligence for predictive purposes." This index avoids the logical fallacy involved in adult mental ages. It is pointed out that the growth function may be a two-parameter curve which would not interfere with the use of sigma values, but would reduce the value of a single parameter statistic. *Harold Bechtoldt*

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Page, J. D. "The Effect of Nursery-School Attendance Upon Subsequent I.Q." *Journal of Psychology*, X (1940), 221-30

Stanford-Binet I.Q.'s of 72 children in kindergarten to the fifth grade who had previously attended nursery school 125 to 525 days were compared with those of adjacent older siblings who had not attended preschool. One hundred children of like age and socio-economic status were also compared with their adjacent older siblings, none of either group having attended preschool. No significant differences in I.Q. could be referred to nursery-school attendance. A slight advantage of younger siblings in both experimental and control groups was explained by age fluctuations in the standardization of the L form of the Stanford-Binet. No relation was found between duration of nursery-school attendance and subsequent I.Q. advantage. The mean I.Q. difference between sibling pairs approximated 10 points. *W. A. Farver*

Powell, N. J. "Check List for Use in Civil Service Objective Test Preparation." *Public Personnel Quarterly*, II (1940-41), 13-6

The article includes a list of questions which have been developed for reviewing civil service objective tests before they are finally used. Its use is intended to "increase the probability that no major basis upon which the test will be appraised has been ignored in the test construction." Points to be checked are listed under the following headings: validity, cost, appearance of test, typography, and administration. A number of questions applying specifically to completion items and multiple-choice items are also listed.

Roff, Merrill. "Linear Dependence in Multiple Correlation Work." *Psychometrika*, V (1940), 295-98

The problem in multiple correlation work of nonsense results attributable to linear dependence of variables, which has been discussed by Ragnar Frisch in relation to economic data, is presented from the standpoint of its significance in psychological research. It is shown that a symmetric corre-

MEASUREMENT ABSTRACTS

lation determinant with unity in the diagonal cells can vanish only when there is a first-order or partial correlation of unity between one pair of the variables. On the basis of this result, it is argued that the problem should be expected to cause less difficulty in the field of psychology than in economics and that psychologists should be able to avoid the pitfall by bringing to bear their knowledge of the variables with which they are working. (Courtesy *Psychometrika*)

Royer, Elmer B. "A Machine Method for Computing the Biserial Correlation Coefficient in Item Validation" *Psychometrika*, VI (1941), 55-9

A method for computing the biserial correlation coefficient with the aid of punch-card equipment is outlined. A numerical example and a work sheet layout are included in the presentation. (Courtesy *Psychometrika*.)

Ryans, David G. *The First Step in Guidance: Self-Appraisal*. New York, Cooperative Test Service. 35 pp. 1941

A report of the 1940 Sophomore testing program in which the following tests were used: Cooperative English Test, Form Q, Cooperative General Culture Test, Form Q, and Cooperative Contemporary Affairs Test, Form 1940.

Sisk, H. L. "A Note on the Comparative Value of the 'True' Index of Studiousness for the Purpose of Prognosis" *Journal of Psychology*, X (1940), 275-78

The scholastic achievement of 585 university freshmen was predicted from Symond's "true" Index of Studiousness and from a battery of tests, composed of aptitude, English, and reading. The latter was found to give a more reliable prediction of first semester grades. *W. A. Varvel*

Stoy, E. G. "Selection of Key-Punch Operators" *Journal of Applied Psychology*, XXIV (1940), 653-54

These are notes on preliminary experimentation in the selection of key-punch operators. Four tests warrant further

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

consideration "an eye-hand coordination test in which letter combinations involving both hands are registered on counters, a test of verbal and spatial memory, a clerical type of test, and an arithmetic test" *W A Varvel*

Swineford, Frances and Holzinger, Karl J "Selected References on Statistics, the Theory of Test Construction, and Factor Analysis" *School Review*, XLVIII (1940), 460-66

Articles covering the year March, 1939, to February, 1940, are presented with brief notes as to the nature of the problem handled in each paper. Twelve articles are given under the heading "Theory and Use of Statistical Methods," 18 under "Problems of Test Construction," and 16 under "Factor Analysis." *Harold Bechtoldt*

Thurstone, L L "A Factorial Study of Visual Gestalt Effects" *Psychometrika*, V (1940), 315-16 (Abstract of a paper read at the September, 1940, meeting of the American Psychological Association)

Toolon, W T "Essential Factors in Test Construction" *Personnel Journal*, XVIV (1940), 204-08

The value of careful "informal examination" of test items before and after statistical treatment is pointed out, and an analysis of the nature of the items and of the errors made is suggested. Factors dealt with include item difficulty, item correlations, closeness of distractors, and the judgment and information of the subject. *Harold Bechtoldt*

Tucker, Ledyard R "A Matrix Multiplier" *Psychometrika*, V (1940), 289-94

A machine to expedite matrix multiplication has been developed by modifying the International Business Machines Corporation scoring machine. The principles and operation of the machine are described, and time and accuracy estimates are indicated. (Courtesy *Psychometrika*)

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Volume I

JULY, 1941

Number 3

A NEW PERFORMANCE TEST FOR YOUNG DEAF CHILDREN	217
<i>Marshall S. Hiskey</i>	
PERFORMANCE TESTING IN PUBLIC PERSONNEL SELECTION	233
<i>Sidney W. Koran</i>	
SOME DATA ON THE KUDER PREFERENCE RECORD	253
<i>Arthur E. Traalen and William C. McCall</i>	
THE RELIABILITY OF RATIO SCORES	269
<i>Lee J. Cronbach</i>	
GUIDING STUDENTS TO BECOME SELF-GUIDING	279
<i>Joseph S. Kopas</i>	
AN ATTEMPT TO MEASURE SCIENTIFIC THINKING	289
<i>Max D. Engelhart and Hugh B. Lewis</i>	
AN EVALUATION OF TECHNIQUES OF MEASURING VISUAL ACUITY AT THE COLLEGE LEVEL	295
<i>Frances Oraland Triggs and Karl E. Sandt</i>	
THE CONCEPT OF SCATTER IN THE LIGHT OF MENTAL TEST THEORY	303
<i>Maurice Lorr and Ralph K. Meister</i>	
MEASUREMENT ABSTRACTS	311
MEASUREMENT NEWS	318

Copyright, 1941, by
SCIENCE RESEARCH ASSOCIATES

PRINTED IN THE UNITED STATES OF AMERICA

A NEW PERFORMANCE TEST FOR YOUNG DEAF CHILDREN

MARSHALL S. HISKEY¹

University of Nebraska

Introduction

THERE has long been a need for a measuring device which would give the teacher of the very young deaf child a valid indication of his learning level at the beginning of his educational career. One can find a considerable number of more or less carefully worked out mental tests which have been used for deaf and hard-of-hearing individuals. However, the degree of help which such tests render the educator or clinician depends upon the number and representativeness of the children upon whom they have been standardized, and also upon the reliability and amount of information about the children which the test makes available. Few tests have been standardized on deaf children or used with such children at the beginning of their school experience.

Instruction, especially at the lower levels, although carried on as a group activity, actually involves considerable individualized work. This individualization is, in many instances, primarily a means of preparing for group instruction. Therefore, if classes are not composed of students of approximately the same level of ability, the

¹The writer wishes to acknowledge his indebtedness to Dr. D. A. Worcester and other staff members of the Department of Educational Psychology and Measurements of the University of Nebraska and to the administrations of the Iowa, Nebraska, Kansas, Missouri, Illinois, Indiana, and Ohio State Schools for the Deaf.

teacher must spend entirely too much of her time working with the slower pupils as individuals. In many instances this is done at the expense of the more capable students and often results in a great waste of time since it is difficult to keep the young deaf child occupied constructively without the direct, and almost constant, guidance and supervision of the teacher. If supplementary measuring devices are valuable in making the school program more effective for the hearing child, then they should be even more valuable with a group who must start with the handicap of deafness.

Difficulties involved in constructing a test for the young deaf and hard-of-hearing. In the selection of items for young deaf children, the special limitations of this group must be kept constantly in mind. The actual testing of deaf children presents problems which are unique. Practically every impression of the test materials gained by the deaf child must be through the sense of sight. All instructions must be given through pantomime. Because of the child's complete lack of language experience, the test items must have an unusual intrinsic attractiveness. In addition to these problems one must devise a sufficient variety of items to sample adequately the abilities of individuals whose range of experiences has been seriously restricted.

To attempt to obtain a rating of the "word fluency" of the child who has been deaf since birth would be futile. Nor does it seem appropriate to include speed tests since it is very difficult to give to the young deaf child the concept of speed.

Based on the observations gained through testing the members of both groups, the writer is of the opinion that deaf subjects are more prone to "jump to conclusions" and to overestimate their abilities or the amount of material which they have grasped, than are hearing subjects. It is necessary to make them take their allotted time for viewing materials before they attempt a response. On the

other hand, the examiner must always be on the alert, lest through some slight change in facial expression he assist the subject in making his response. The deaf or hard-of-hearing child is continuously seeking visual clues and an "arched eyebrow" or the "flicker of an eyelash" may speak volumes to him.

The writer has made no attempt to compare the intellectual development of deaf and hard-of-hearing children with that of hearing children. The deaf child's training probably will never be identical with that of the hearing child. The writer is of the opinion that the question of primary importance is not, "How does the deaf child rank in comparison with the hearing child?", but rather, "How does the deaf child rank in comparison with other deaf children of his chronological age?"

Development and Standardization of the Scale

Preliminary study of deaf and hard-of-hearing children in school. In order to obtain a more adequate understanding of the group, the writer made an intensive study of deaf children as they actually went about their school work.

For a period of more than four months the writer spent three days every two weeks with these pupils in a residence school for the deaf. Not only did he visit them at their class work but he lived with them at the school and associated with them on the playground, in the gymnasium, and elsewhere. A complete record was made of the activities which took place in the classroom and also of those of an extra-curricular nature. This type of study yielded a multitude of suggestions which were of the utmost importance in the construction of the scale.

The selection and construction of test items. Every item of the scale was considered in light of the following criteria: (1) Was the item similar to the task, or tasks, which the young deaf child did in school? (2) Was it the type of item which could be included in a non-verbal test? (3) Could the item be presented in such a way that

directions could be given through simple pantomime? (4) Was it the type of item which experience had shown to yield high correlation with acceptable criteria of intelligence or learning ability? (5) Could the item be constructed and presented in such a way that the child could give a definite response, thus making the scoring objective and easily done? (6) Would the item be appealing or attractive to the subject? (7) Could the item be scored without the score being based on time? (8) Did the difficulty of the item appear to be within the age range of the standardizing group? (9) Did the item seem likely to show a high discriminative capacity?

In many instances, in order to meet all the above criteria, it was necessary to devise special methods of constructing or assembling the parts of an item. The preliminary scale was composed of 18 different types of items with a total of 204 items.

The use of the preliminary scale This scale was given to seventy-three pupils of the Iowa School for the Deaf, whose ages ranged from three years ten months to nine years eight months. Owing to the length of the scale, it was divided into two parts and half of the group was given Part A first and the other half of the group was given Part B first. The two parts were given not less than one day nor more than one week apart. In several instances items were scored in detail, thus permitting a later rescore on a different basis.

After members of this tryout group were tested, an item analysis was made and curves of the percentage passing each successive chronological age were plotted. This was done for each of the 204 individual items of the scale. The steepness of these curves afforded a graphic indication of the validity of the items. The items which appeared to function the most satisfactorily and to most nearly approximate the criteria were retained. The criteria used were (1) validity (based on the percentage passing from one age to the next), (2) ease of adminis-

PERFORMANCE TEST FOR DEAF CHILDREN

tering, (3) ease and objectivity of scoring, (4) attractiveness or interest to the subject, (5) variety; and (6) time of administering. When the sifting process was completed, 11 types of tests were retained, including a total of 124 individual items.

The test items. A brief description of the items may make later discussions more meaningful. The types are as follows:

- 1 *Memory for Colored Objects*—Two sets of eight colored sticks each, one set for the examiner and one set for the subject. The examiner presents from one to five of the sticks from his group and then removes them and the subject must select the corresponding sticks from his set from memory.
- 2 *Bead Stringing*—At the lower levels scoring is based on the number of beads strung during a two-minute period. The intermediate level demands the correct copying of bead patterns, while at the upper level the subject is rated on his ability to reproduce patterns from memory.
- 3 *Pictorial Associations*—This includes 12 series of pictures. In each series two pictures are mounted side by side and a recess is left for the insertion of the third picture which is associated with the first two. This third picture must be selected from a group of four unmounted pictures. (There are four unmounted pictures for each series.)
- 4 *Block Patterns*—A set of eight drawings of block patterns and 16 blocks. The patterns are arranged in order of difficulty and the subject must construct the pattern shown in the drawing.
- 5 *Memory for Digits*—Two sets of nine numbers each. The examiner presents a number series and the subject must reproduce it from memory.
- 6 *Completion of Drawings*—A series of 15 pictures, each with a part missing. The subject must draw the missing part and thus complete the picture.
- 7 *Pictorial Identification*—Six series of mounted pictures. Each series has five pictures of a similar nature which are mounted side by side. Four individual pictures which are duplicates of the mounted pictures must be correctly identified by matching them with the corresponding mounted picture.
- 8 *Paper Folding*—Six-inch squares of paper which must be folded by the subject to duplicate (seven) patterns.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

- 9 Visual Attention Span—Several series of pictures (varying from one to six pictures each) and 15 individual pictures. The subject is shown a picture series and he must use the individual pictures to reproduce the presented series from memory.
- 10 Puzzle Blocks—Eight sets of variously shaped pieces of wood. Each set can be put together to form a block.
- 11 Pictorial Analogies—Ten series of pictures with three pictures in each series mounted and four pictures to use as choices. The first, second, and third pictures of the analogy are mounted and a recess is left for the insertion of the fourth picture which completes the analogy. The subject must select the latter picture from among the four available choices.

Use of the provisional scale. In addition to the work done with the pupils of the Iowa school, the test was administered in the state schools for the deaf in Nebraska, Kansas, Missouri, Illinois, Indiana, and Ohio, as well as to the members of the Lincoln, Nebraska, Day School. All students, except a few who were ill, who were under 10 or who had had their tenth birthday within 15 days of the examination date were tested. The test was administered to 466 individuals. The standardizing group is limited in numbers at the age of four and below, since most schools do not accept children until they are five or six years of age.

Derivation of the final scale. To save time and to guarantee greater accuracy in the statistical data on which the final selection of items would be based, Hollerith techniques were used. By means of the Hollerith sorter and counter it was possible to determine quickly the number of individuals who were successful on each item in successive ages throughout the range and thus to plot for each item the curves of percentage passing. Items were selected chiefly on the basis of discriminative ability, this judgment being based on the increase in percentage passing from one age to the next. The items in each group were next arranged in order of difficulty, this order being based on the percentage of the total group passing each individual item.

PERFORMANCE TEST FOR DEAF CHILDREN

To develop the table of norms, curves were plotted showing for each age group the percentages making each possible total score for each group of items. The score necessary for passing each item at a certain age level was considered to be that score which was made by approximately 70 per cent of the particular group. In all instances the percentages were plotted, the curves were smoothed, and the ends were extended to obtain what might be termed "projected norms" at the extremes. This smoothing of the curves of percentages gives a somewhat truer indication of the ability level of the four-year-old group.

To determine who should compose the four-year-old group, or the five-year-old group, etc., it was decided to classify all individuals as four whose ages were between three years six months and four years five months and as five those who were between four years six months and five years five months, and so on. In no instance does the mean chronological age of the standardizing group deviate more than one month from the desired or true mean.

The unit of measurement Perhaps the most common method of interpreting scores on a scale such as this one is the familiar Binet type mental age. This is the method of using age norms and the amount of mental development in a year as the unit of measurement. Age norms are established for raw scores and are converted into, or interpreted as, mental ages. This age-type score, representing the amount of development up to date, has much greater meaning to the layman than does the "standard score" or the "percentile score" and for that reason the age norm has been used in this scale. However, the term "mental age" has not been used because the M A would undoubtedly suggest a Binet Mental Age which in turn would suggest the corresponding M A of the hearing child and thus lead to false comparisons. For this reason and because of the fact that the test items have been selected, in many instances, because of their similarity to

the abilities which the deaf child must exhibit in school, the term "Learning Age" is used instead

An L A. of 5-0 simply means that, according to the results of this test, the child is able to do those tasks which the average deaf child of five years is able to do, or, that he should be able to solve problems with the same average efficiency as the average *deaf* five-year old

It is recommended that in the interpretation of test results, the learning age be used instead of the learning quotient (L.Q., derived by dividing the L A. by the C A., similar to the I Q.) Until more conclusive evidence regarding the respective influence of environment and heredity on the mental development of the child, resulting from more carefully controlled experiments, is produced, one must proceed cautiously to insure that he is not closing the door of opportunity to any child. If there is a reasonable question as to whether the hearing child can be improved through a stimulating program of training, is it not likely that this question will assume even larger proportions in the case of the deaf child?

Statistical Analysis of Test Data

The accuracy of any test is dependent, not only upon test items employed, but also upon the number of individuals examined, the representativeness of the group, the accuracy with which the test has been scored, the derivation of accurate and meaningful norms, and various statistical applications which are used for the purpose of checking, or for interpretation. An additional and more detailed statistical treatment of the data will be made at some later date. Such topics as sex differences, effects of schooling, relation of score to degree of hearing loss, resemblances of score to teacher judgment of ability, and the results of a factorial analysis of the test items, are among those so reserved.

Adequacy of the standardization. Perhaps the main criterion for the standardizing of any test is the selection of representative populations at each age. The method

PERFORMANCE TEST FOR DEAF CHILDREN

employed in meeting this problem has been described briefly above, i.e., the testing of all available pupils (within the desired age range) in a rather widely scattered group of state schools for the deaf. However, to check the adequacy of the sampling of cases, a table of percentages of scores for each item was made which did not include the students of the Indiana school and the Ohio school. From this list of percentages, a table of norms was derived. These norms were then compared with the norms derived from the total group. In 89 per cent of the cases, the norms were found to be identically located and in the remaining 11 per cent of the cases they varied not more than six months. This would indicate that the sampling was probably sufficient for determining relatively stable norms.

TABLE 1
YEARS-IN-SCHOOL DISTRIBUTION BY AGES AND
A COMPARISON OF THE MEAN C.A.'s AND THE MEAN L.A.'s
FOR THE STANDARDIZING GROUP

Age	Years in School							Total	Mean C.A.	Mean L.A.
	0	1	2	3	4	5	6			
4	9	1						10	4-1	4-4.8
5	33	9						42	5-0.7	5-1.8
6	39	16	4	1				60	6-0.3	6-3.5
7	22	42	15	4	1			84	7-0	7-2.3
8	11	31	27	14	4			87	7-11.4	8-0.7
9	6	29	43	31	5	3		117	8-11.7	9-2.9
9-9	5	9	15	17	12	6	2	66	9-9	9-6.5
TOTAL	125	137	104	67	22	9	2	466		

Table 1 gives the number of individuals tested at each chronological age level. The small number of cases in the lowest age group means that the norms will be less reliable at the age of four. This table also shows the mean chronological age and the mean learning age for each of the age groups. In no instance does the mean chronological age differ more than one month from the desired chronological age. The mean learning ages likewise correspond closely to the mean chronological age. At each age level, except one, the mean L.A. is slightly higher than the mean C.A. It is felt that this is a desirable fea-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 2
PER CENT OF EACH AGE GROUP

DEAD STRINGING										BLOCK BUILDING							
Total per 2 Min					Total Patterns					Total Score							
Age	7-8	9-10	11-12	13-14	I	II	III	IV	V	1	2	3	4	5	6	7	8
4	100	90	40	30	40	10				100	100	30	10	10			
5	100	91	79	55	60	29	12			100	100	81	40	17	2		
6	100	97	94	87	92	47	20	5		100	100	98	82	52	27	8	3
7	100	99	99	97	96	87	45	12	2	100	99	99	90	73	48	24	6
8	100	100	100	100	100	95	67	32	9	100	100	99	94	85	69	44	11
9	100	100	100	100	100	99	88	57	19	100	100	99	98	95	83	66	39
9-9	100	100	100	100	100	100	89	70	21	100	100	100	100	98	95	80	50

PICTORIAL ASSOCIATIONS													PAPER FOLDING						
Age	Total Score												Total Score						
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7
4	100	100	100	90	60	20	20						100	100	100	40	10		
5	100	100	100	93	79	40	26	12	2				100	100	100	88	71	26	
6	100	100	100	97	90	80	62	43	20	8	2		100	100	100	97	90	72	18
7	100	100	99	99	95	89	83	67	54	23	6	1	100	100	100	98	95	89	61
8	100	100	100	99	99	97	94	85	67	51	18	6	100	100	100	99	99	96	69
9	100	100	100	100	99	99	97	96	82	73	39	12	100	100	100	100	100	99	94
9-9	100	100	100	100	100	100	100	100	91	82	58	14	100	100	100	100	100	100	95

MEMORY FOR COLORED OBJECTS													MEMORY FOR						
Total Score													Part A Totals			Not in Order			
Age	5	6	7-8	9	10	11	12	13	14	15	16	17	3	4	5	B	C	D	E
4	100		90	40	20	10							100	80	10	10			
5	100		98	62	26	14	2	2					100	88	45	64	10	10	10
6	100		98	90	70	52	27	15	5	2			100	95	87	97	65	10	10
7	100		99	98	89	77	54	31	13	6	1		100	99	97	98	83	38	15
8	100	100		99	93	84	71	60	40	22	11	3	100	100	100	100	94	59	17
9	100	100	100		97	91	85	73	56	43	27	5	100	100	100	99	94	79	37
9-9	100	100	100	100	100	100	95	80	62	48	29	17	100	100	100	100	98	83	52

ture If there were an inadequate sampling of subjects, it would likely be of the group with limited ability, since the mentally less advanced are less likely to have entered school at a reasonably early age than are those mentally advanced. The test ceiling is not high and this may be responsible for the fact that the group at the upper end of the range has a mean L A slightly below their mean C A. In only two instances do the two means deviate by as much as three months—the greatest deviation being 3.8 months at the four-year level.

PERFORMANCE TEST FOR DEAF CHILDREN

MAKING EACH SCORE ON EACH TYPE OF ITEM

PICTORIAL IDENTIFICATION													VISUAL ATTENTION SPAN						
Total Score													Total Score						
1	2	3-4	5-6	7-8	9-10	11-12	13-14	15-16	17-18	19	20	21-22	23-24	1	2	3	4	5	6
100	100	100	100	100	90	80	50	40	10					100	70	20	10		
100	100	100	100	100	98	95	93	79	60	33	10	5		100	73	29	2		
100	100	100	100	100	98	93	80	65	48	29	23			100	95	72	35	12	2
100	100	100	100	100	99	99	98	96	94	85	75	60		100	99	90	57	21	3
100	100	100	100	100	100	100	100	99	98	98	95	87		100	100	98	77	36	11
100	100	100	100	100	100	99	99	98	98	98	98	93		100	100	98	85	51	26
100	100	100	100	100	100	100	100	100	100	100	100	98		100	100	100	88	70	32

PUZZLE BLOCKS							PICTORIAL ANALOGIES									
Total Score							Total Score									
1	2	3	4	5	6	7	1	2	3	4	5	6	7	8	9	10
90	30						10	10	10	10						
100	95	55	12				100	76	71	43	14	5	2			
100	98	80	52	12			100	100	95	80	57	30	12	2	2	2
100	98	87	73	38	5		100	99	96	89	79	52	27	10	1	
100	100	99	89	64	22	2	100	100	100	98	87	63	43	28	2	
100	100	100	97	77	54	14	100	100	100	98	94	91	76	56	26	7
100	100	100	100	89	62	21	100	100	100	100	98	95	83	64	30	8

DIGITS				COMPLETION OF DRAWINGS													
In Order				Total Score													
B	C	D	E	1	2	3	4	5	6	7	8	9	10	11	12	13	14
				20	10	10											
50	2			67	44	36	19	17	12	7	2	2	2				
88	33	3		98	93	87	77	70	55	38	25	20	10	7	2		
91	55	13		98	96	94	87	85	77	73	67	52	43	30	19	7	
98	80	32	6	100	100	99	98	95	95	94	92	88	74	53	34	16	7
98	83	53	24	100	100	100	100	99	99	97	96	88	86	79	71	47	16
100	97	73	32	100	100	100	100	100	100	100	100	100	97	94	83	61	23

Table 2 gives the percentages of subjects in each age group making each possible total score for each type of test item. These per cents were the ones used in plotting the curves of per cents. These curves were smoothed and the score which revealed approximately 70 per cent of success was taken as the score of the average person of that chronological age. It was then entered in the table of norms under that learning age. The learning ages given below 4-0 and above 10-0 are the results of an extension process and, as has been mentioned before, are not so

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

reliable as are those within the age range of the subjects examined

Validity The very methods by which the test items have been selected and retained are evidence of their validity. It will be recalled that the items were selected according to rather definite criteria and that after they had been given they were subjected to a rigorous item analysis. Thus the chief criteria for validity were (1) selection—through critical analysis and adherence to criteria, and (2) increase in the percentage passing from one age to the next. In the present scale, it was impossible to determine validity through correlations with other test scores inasmuch as there is no existing test which would have been an acceptable criterion. In the absence of the needed criterion, correlations were computed between the score on the entire scale (the score on the entire scale is the median learning age of the learning ages obtained on the several parts of the scale) and the score on each group of items. The correlation setup is seemingly a spurious one since a part of the test has been correlated with the whole test which includes this part. As the score on the entire scale is the median score of the parts of the scale, however, each part has an approximately equal share in producing this total or final score and this in turn lessens or eliminates the possibility of

TABLE 3
CORRELATIONS BETWEEN THE LEARNING AGE OBTAINED ON ONE
SECTION OF THE TEST AND THE MEDIAN LEARNING AGE
OBTAINED ON THE ENTIRE TEST

	Group I (Age 4 to 7)	Group II (Age 8 to 10)
1 Memory for Colored Objects	804	740
2 Bead Stringing	812	729
3 Pictorial Associations	643	693
4 Block Building	797	718
5 Memory for Digits	755	773
6 Completion of Drawings		702
7 Pictorial Identification	730	
8 Paper Folding	843	
9 Visual Attention Span	637	629
10 Puzzle Blocks		734
11 Pictorial Analogies		742

the obtained correlations being spuriously high. Since the correlations between the learning age obtained on each group of items and the median learning age on the entire scale are within the range of from .629 to .843, they are evidence of high internal consistency and thus, perhaps, of high item validity.

The abbreviated scale. To determine whether a dependable short scale could be assembled, the five types of items which showed the highest correlation with the median learning age for the entire scale were selected to form the abbreviated scale. Since some of the groups of items do not function over the entire age range, correlations were derived separately for two groups. Group I was composed of all members of the standardizing group who were seven years or under, and Group II, those who were from eight to 10 years of age. For Group I, correlations with the total scale were obtained for all groups of items except those which do not function at the lower levels, and for Group II, correlations with the total scale were obtained for all groups of items except those which do not function at the higher levels. The best booklets were rescored on the basis of these abbreviated scales and correlations were found between the median learning ages obtained from the abbreviated scales and the original scale. The correlation for Group I was .944 and for Group II .936. Thus, when time limitations make it necessary, the short forms may be used with a considerable degree of confidence. These abbreviated forms can be given in approximately 30 minutes.

Although it is recommended that the learning age be used in preference to the learning quotient ($LQ = LA / C.A.$), in order to make the study more complete and significant the writer has made a study of the LQ's of the standardizing group. Table 4 shows that the mean learning quotients derived from the standardizing group closely approximate the desired mean of 100. The greatest deviation is at age four and is probably due to the

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

small number of cases and to the fact that they are a somewhat select group

TABLE 4

THE MEAN LEARNING QUOTIENT, RANGE, STANDARD DEVIATION OF THE LQ'S, AND STANDARD ERROR OF THE MEAN FOR EACH AGE LEVEL OF THE STANDARDIZATION GROUP

Age	Mean L A	Range of L Q 's	σ L Q	σ M
4	108.5	94-127	10.909	3.4500
5	102.7	80-124	10.518	1.6228
6	104.4	65-139	14.470	1.8681
7	103.4	43-132	15.300	1.5912
8	101.7	65-137	15.135	1.6226
9	104.0	55-134	14.361	1.3277
9-9	99.8	73-120	11.410	1.4040

The mean L.Q. at each age level except the upper group (9-6 to 10-0) is slightly above 100. As has been mentioned before, this is probably a desirable feature. The lower mean of the upper group is apparently the result of the limited test ceiling. The standard deviations at each age agree closely, except at the two extremes where the attenuating factors before mentioned have influenced them. In general, the standard deviation of the means is approximately 1.6 (disregarding the four-year group).

Every effort has been made to make the test usable and yet have the mechanics as simple as possible. The record blank (Table 5) has been no exception and has been patterned after the one devised by Hildreth and Pintner. The record blank is in reality a table of norms and the various scores are checked on the blank and the median score is calculated. The items of the abbreviated scales also are indicated on the blank.

The test items not only are attractive to young deaf children but also they have a rather high discriminative value. The scale is not difficult to administer or score and since it is weighted heavily with tasks similar to those which the deaf child must do in the early years of his educational career, it should be extremely valuable for gaining a better understanding of the abilities of the younger

PERFORMANCE TEST FOR DEAF CHILDREN

TABLE 5
RECORD BLANK FOR
NON-VERBAL TEST OF LEARNING APTITUDE
(Especially adapted for young deaf children)

NAME	SEX	SCHOOL	GRADE
AGE	BIRTH DATE	L A.	EXAMINER
DATE			
LEARNING AGE	3-0 3-6 4-0 4-6 5-0 5-6 6-0 6-6 7-0 7-6 8-0 8-6 9-0 9-6 10-0 10-6 11-0 11-6		
MEMORY FOR	3- 5- 7-		
†*COLORED OBJECTS	4 6 8	9 10 11	12 13 14 15 16
BEAD	5- 7- 9- 11- 13-		
†*STRINGING	6 8 10 12 14	I II III IV V	
PICTORIAL ASSOCIATIONS	3 4 5 6	7 8 9 10 11 12	
BLOCK			
*PATTERNS	1 2 3 4 5- 6 7 8 9 10 11 12 13		
MEMORY			
†FOR DIGITS	1 2 3 4 5 6 7 8 9 10 11 12 13		
COMPLETION			
OF DRAWINGS			
PICTORIAL	9- 11- 13- 15- 17- 19- 21- 23-		
*IDENTIFICATION	10 12 14 16 18 20 22 24		
PAPER			
*FOLDING	1 2 3 4 5 6 7		
VISUAL			
ATTENTION SPAN	1 2 3 4 5 6		
PUZZLE			
†BLOCKS	1 2 3 4 5 6		
PICTORIAL			
†ANALOGIES	2 3 4 5 6 7 8 9		

*Abbreviated scale for ages 3 to 7
†Abbreviated scale for ages 8 to 10

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

deaf children. This is not intended to imply that the inexperienced person could give the test satisfactorily. The person who is unfamiliar with individual testing techniques would have considerable difficulty unless he underwent a period of training or practice with this scale. It is quite conceivable that the person who has had some experience in individual testing and who has some knowledge of deaf children could, after a period of training in which he gave six or eight practice tests, administer the scale quite satisfactorily.

PERFORMANCE TESTING IN PUBLIC PERSONNEL SELECTION

PART I

SIDNEY W. KORAN¹

Employment Board, Pennsylvania Department of Public Assistance

Introduction

IT IS an interesting fact that although the use of performance tests in the selection of public personnel enjoys not only the general endorsement of personnel technicians but the enthusiastic and unsolicited support of the public as well, there is probably no other aspect of the examination process at present more completely neglected by the majority of merit system agencies.

Probably all jurisdictions employ performance tests in the selection of typists and stenographers, and the general practice is to convert the ratings on these tests into quantitative terms capable of combination with scores achieved in other portions of the examination battery. Beyond that, however, the performance testing of most agencies seems seldom to go beyond the administration of qualifying tests to a sufficient number of individuals at the top of certain registers to satisfy immediate certification requirements. Except for the case of tests of typing

¹The author desires to express his appreciation to the following individuals: Mrs. Ruth Glenn Pennell, and Mr. Robert Hall Crug, members of the Employment Board; Miss Hilda P. Thompson, the Board's Executive Director; Dr. C. H. Smeltzer, the Board's Technical Consultant; Miss Kathleen Oyster, Traffic Representative of the Bell Telephone Company's Harrisburg office; Mr. Andrew S. Hay, Service Supervisor of the IBM Harrisburg office; Mr. Bernard Gehring of the Multigraph Sales Agency in Harrisburg; and Miss Alice I. Thompson, of the Penn State Alumni Association.

and stenography, the technique of using the performance test as a *major part* of the test battery—that is, as a factor which may decidedly influence an examinee's relative standing on the eligibility register—appears to have been almost completely ignored

There are, of course, various reasons why this situation exists. Associated with the technical difficulties inherent in the construction and administration of the performance test—difficulties which, incidentally, are frequently not nearly so “insurmountable” as they at first appear—may be the factors of cost and already overburdened technical staffs. In addition, newly created agencies frequently face time deadlines which all but preclude their going beyond the commonly accepted minimum selection elements, namely the use of minimum requirements, a written test, an evaluation of training and experience, and, for certain positions, an oral interview. In addition to these factors, however, and probably overshadowing them in effect, must be mentioned two others: general inertia and, probably very closely related, an uncritical adherence to time-honored examination patterns considered satisfactory in selecting persons for jobs not requiring the possession of manual skills.

Considerable progress has already been made in the development of performance tests for the selection of typists and stenographers. Since their use is so widespread, no further attention will be devoted to them in this discussion beyond pointing out that, despite their popularity, much necessary work still remains to be done toward their improvement, especially in the development of (1) suitable standards of performance, (2) satisfactory scoring procedures, and (3) improved standardized techniques for administering the stenographic portion of the test.²

²The Chicago Park District's use of phonographic recordings and the novel experiments of the Buffalo Municipal Civil Service Commission and the Arizona Unemployment Compensation Merit System Council with radio broadcasting are examples of approaches to the problem of minimizing or eliminating the undesirable effects of varying dictation speeds and other factors which characterize the use of numerous proctors.

Techniques have also been developed for measuring performance in other jobs such as chauffeur and in certain skilled trades. Many companies test applicants for chauffeur positions and most state motor vehicle bureaus give qualifying driving tests to applicants for operator licenses. The latter are ordinarily quite informally conducted, but Viteles has described "a trade test of driving skill"³ which could quite readily be adapted to merit system use.⁴

The New York City Civil Service Commission's excellent pioneer work in developing tests for such skilled trades positions as welder, machinist, electrician, locksmith, lineman, and carpenter is quite well known.⁵ Recently the State Technical Advisory Service of the Social Security Board began work on standardizing a performance test for Key Punch Operators.

In general, however, the use of the performance test as a measuring instrument designed to serve not only as a qualifying hurdle, but also as an important factor in determining the examinee's relative standing on the eligibility register has received much less attention than it deserves. The dearth of literature on the subject attests to this and probably contributes to the widely held feeling that performance tests capable of producing quantitative ratings are somehow exceptionally difficult to prepare and impractical to administer.

It is hoped that this presentation will illustrate some of the possibilities by describing four actual tests used successfully by a medium-sized merit system agency, the Employment Board of the Pennsylvania Department of Public Assistance. Small jurisdictions may be able to use some of the material with few or no changes. Larger agencies, especially those with adequate technical staffs,

³M. S. Viteles, *Industrial Psychology* (New York: W. W. Norton Company, 1932), 221-24.

⁴The Los Angeles City Civil Service Commission has developed tests of this kind for the positions of Auto Fireman, Ambulance Driver, and Motor Truck Driver.

⁵Fifty-sixth Annual Report—1939 And First Half Of 1940, Civil Service Commission, City of New York.

will probably want to develop their own examinations. Even the latter, however, if their experience with this type of test has been limited, may find it helpful to consider another agency's approach.

Building the Performance Test

The initial steps to be followed in constructing the performance test do not differ in any important respect from those basic to the construction of written tests. In both, the starting point is careful analysis of the job for which the test is to be designed. In constructing a performance test the analysis *must include* actual on-the-job observations of both the equipment and the persons doing the work. If the test constructor is not himself a competent operator of the machine, it will not suffice for him to confine himself merely to study of the printed job specifications and technical literature and to conversations with experts and workers. Such an approach is inadequate even in the construction of written tests, where it is all too often the rule rather than the exception, as the only preparation to designing a performance test it can produce very unfortunate results.

Every one of the foregoing steps—study of the job specifications as part of the agency's classification plan, study of technical literature available on the equipment and on its operation, and conferences with skilled workers, supervisors, and acknowledged experts in the field—has its place in the procedure. That place is as a *supplement* to a first-hand acquaintance with the job itself. The professional test constructor must analyze the job sufficiently thoroughly to permit himself to identify the skills involved and to determine their relationship to one another and to the whole, and he must discover those individual differences which will provide him with essential clues to types of test items likely to prove valid in differentiating among various levels of performance ability.

Here it may be worth pointing out that there is per-

PERFORMANCE TESTING IN PUBLIC PERSONNEL

haps no other phase of the examination program in which the personnel technician is less likely to turn out a satisfactory job unless he consults with specialists who know the practical and technical aspects of the job to be tested. Both specialists—the personnel technician *and* the expert in the occupational field under consideration—bring to the task certain information and knowledge of techniques which need to be reconciled toward a common end, that of producing a valid measuring instrument capable of fulfilling the numerous practical considerations which public agencies cannot afford to forget. The test constructor will want to find an expert who knows the job and who is sufficiently progressive, adaptable, and interested in the problems of personnel selection to be cooperative and sympathetic. The length of time required to orient such a co-worker in the problems of testing will not be great, and the effort will pay big dividends in the form of a smooth working relationship, a valid measuring instrument, and a strong ally in the event of later criticism.

It should be kept in mind that the performance test should (1) be sufficiently long to include an adequate sample of the differentiating essentials of the job, (2) be as inexpensive and easy to administer as possible, (3) minimize possible differences in achievement resulting from lack of immediate familiarity with the particular model of equipment on which it is given, (4) appear sufficiently practical and comprehensive to create a favorable impression among those who do not qualify as well as among those who do, (5) be capable of uniform administration to all candidates, (6) be objectively scored and produce quantitative ratings.

Setting the Passing Point

Since one of the functions of the performance test is to eliminate candidates who do not demonstrate adequate ability to operate the equipment, passing points must be

established with considerable care. Fortunately, this problem can usually be approached much more directly when performance tests are involved than it can with written tests. Practical considerations ordinarily make it impossible for most jurisdictions to employ standardized written tests or to develop satisfactory norms for the tests they construct. The usual practice, therefore, in agencies not bound by restrictive "70 per cent passing" legislation is to permit such factors as the following to influence the location of written test passing points: the number of examinees, the number of openings likely to occur during the life of the register, the general caliber of the competing group, whether or not the examination battery includes such other hurdles as a performance test or an oral interview, and previous certification experiences concerning the ratio of refusals to acceptances.

In establishing the qualifying point for a performance test, on the other hand, the principal criterion must be an affirmative answer to the question, "Can the examinee perform the task well enough to meet the employer's minimum standards?" Production records are ordinarily available on types of work sufficiently similar to those sampled by the test to serve as the basis for setting the elimination point. Where such records are not available or are not in usable form, they can generally be obtained quite easily, and profitably, too, during the test technician's study of the job.

While such data should usually serve as the principal basis for establishing the qualifying grade, they ought not to be the sole consideration. Some of the other factors, for example, which it is frequently important to note are (1) the level of ability of the agency's employees as compared with that of other persons doing the same kind of work, (2) the immediate, and possible future, condition of the labor market in the specific field under consideration and in related fields, and (3) the possible effects of nervousness, atypicality of the test situation, and other factors

likely to be present and to lower the validity and reliability of the examination

Four Performance Tests

The remainder of this presentation will be devoted to describing, with a minimum of discussion, some of the forms and procedures developed in connection with performance tests for the following four kinds of jobs: Telephone Operator, Graphotype-Addressograph Operator, Tabulating Machine Operator, and Duplicating Machine Operator

In considering the material, the reader should keep the following facts in mind

- 1 Each of the tests was designed for examinees who had "passed" a previous hurdle, that of scoring above the 60th percentile in the combination of their written test score and training-experience rating
- 2 The law under which the examining agency operates forbids the establishment of any kind of minimum training and experience requirements
- 3 The operating agency prefers to make no provision for training new employees for these positions and requires that a new appointee be able to perform the duties of the position almost immediately
- 4 Each of the tests was designed to serve as a qualifying examination capable of weeding out individuals lacking in sufficient operating ability to produce satisfactory work, and as a measuring instrument capable of producing quantitative ratings of relative ability to perform the work
- 5 The validity of none of the tests has been determined through the use of statistical procedures. For the present, their only claim to validity is based on the fact that (1) experts in the fields covered by the tests state that they measure what they purport to measure, and (2) employees certified from registers established on the basis of these tests have proved more uniformly satisfactory to their employers than those certified from eligibility lists set up from examination batteries which did not include performance tests⁶

⁶Studies of reliability and of the relation between performance test scores and service ratings, written test scores, training scores, and experience scores are now under way

The Test for Telephone Operators

This test was designed to be administered to examinees in the 20 counties of the state in which vacancies existed for the position of telephone operator in either the main, or a regional, office of a local Board of Assistance. The counties were unusually widely scattered geographically. The dial system was in use in 60 per cent of the 20 counties and the manual system in the remainder, but the PBX boards were all of the cord-type. The switchboards in 90 per cent of the offices were connected to four or more trunk lines, the largest two had 12 and 15 trunk lines respectively. The number of extension stations varied from 11 to 66, but 60 per cent had 20 or more.

The preliminary survey of the physical facilities available in each of the 20 counties for which eligibility registers were to be established was accomplished by means of a letter in which a questionnaire containing the following questions was enclosed:

1. Is your PBX switchboard of the cord type?
2. How many city trunk lines do you have?
3. Is a dial part of the equipment of your switchboard?
4. Is the entire city in which your office is located equipped with dial telephones?
5. How many extension stations do you have?
6. Is there an office near your switchboard in which there are two separate extension lines (not two extensions of the same line)?
7. If so, is the office in which these two lines are located within hearing distance of the ring of a third line?
8. Would you be willing to permit us to use your switchboard for the performance test if the test is scheduled on a Saturday afternoon or on a week day evening when there are few or no business calls likely to interfere?

Twelve examination centers were established. The factors which dictated their selection were: (1) the type of equipment available in each county for which a register was to be established, (2) the relative proximity of counties with similar equipment, (3) the distance each examinee would be required to travel, and (4) the cost of administration.

PERFORMANCE TESTING IN PUBLIC PERSONNEL

The test itself consisted of a series of 13 operating situations designed to determine the examinee's ability to service incoming, outgoing, and extension calls and transfers of incoming and outgoing calls, all under as nearly normal operating conditions as were practically possible to achieve. The minimum equipment necessary for the administration of the test was a cord-type switchboard having four trunk lines and five extensions. Two forms of the test were required: one, Form D, for operators of PBX installations in communities where the dial system was in use, the other, Form M, for manually operated installations.

The test was administered by two persons (designated, in the test, as Mr. Albert and Mr. Brown), one of whom was required to be well acquainted with the procedure and to have had some practice in its administration. The two examiners used separate extension telephones but were situated within earshot of each other and within hearing distance of a third extension telephone.

The administration, recording, and scoring of the test were facilitated by the development of a combined "cue sheet" and rating form designed to serve the four-fold function of (1) indicating the sequence of operations so that each examiner would know what his task was at every stage of the test, (2) listing the phrases to be repeated verbatim by both examiner and examinee, (3) enumerating the items on which the examinee was to be rated, and (4) providing spaces for the examiners' ratings and comments. Each examiner was provided with a copy of this form and, as Mr. Albert or Mr. Brown, was required to originate, maintain, and terminate the calls assigned to him and to rate the examinee on each phase of every call coming to his attention.

As an aid in orienting the examinee to the test situation she⁷ was provided with an Instruction Sheet (Ex-

⁷The feminine pronoun is used because all examinees for the telephone operator performance test were women.

Exhibit A) which set forth the general nature of the test she was about to take and listed a few simple instructions such as any experienced operator would need to be given on starting a new job. When the examinee had been permitted ample time to read the Instructions, she was assigned to the switchboard. Several minutes, if necessary, were then allowed to permit the examinee to familiarize herself with any aspects of the board which were strange to her and to note the location of the jacks and names mentioned in the Instructions. When the examinee was ready to begin, the receptionist told her to ring Mr Brown's extension and to read her identification number to him from her admittance slip. This operation, as well as the first call placed by the examiner, was intended to help "break the ice" and did not enter into the determination of the examinee's score.

Exhibit B is a reproduction of the test administered to examinees required to operate switchboards in dial-equipped communities. While the calls comprising the manual form of the test were similar in number and complexity to those included in the dial form, different cue and rating sheets were required because several of the operations (and, consequently, the points to be rated) were not the same for both systems.

Some idea of the variety of realistic operating situations existing during the administration of the test—despite the fact that all of the calls were originated by only two persons—may be gathered from an examination of some of the calls the operator is required to handle. In call No. 4 (see Exhibit B), the operator connects Mr Albert's extension to a city line. A few moments later, the operator is telling the person whose call has come in on a trunk line that Mr. Carson's extension is busy (call No. 5). To maintain the connections required at this stage of the test the operator had to put up seven cords. In the four calls immediately following, the operator was required to perform these tasks:

PERFORMANCE TESTING IN PUBLIC PERSONNEL

Call No 6

answer Mr Albert's extension,
transfer the incoming call from Mr Carson's extension to Mr Albert's,
take down the connection from one of the trunk lines and from Mr Albert's and Mr Brown's extensions

Call No 7

answer Mr Brown's extension,
connect Mr Brown's extension to a city line so that Mr Brown may dial his number through the central exchange

Call No 8

answer an incoming call,
inform the person calling that Mr Brown's line is busy,
hold the incoming call until Mr Brown's line is no longer busy

Call No 9

answer Mr Albert's extension,
transfer the outgoing call from Mr Brown's extension to Mr Albert's extension,
take down the connections from one of the trunk lines and from Mr Albert's and Mr Brown's extensions

The scoring procedure was designed (1) to permit the immediate elimination of candidates whose performance fell below certain established minimum standards, and (2) to produce quantitative ratings reflecting the relative operating ability of the examinees who satisfied these minimum standards. Because both the level of difficulty of the duties and the relative ability of candidates to perform the duties varied in direct relation to the size of the county in which the jobs occurred, minimum standards (based on the 12 calls comprising the test) were set on a class-county basis as follows:

Class II Counties	10 completed calls
Class III Counties	9 completed calls
Class IV Counties	8 completed calls

For the purpose of applying the minimum requirements represented by these criteria, a call was considered "completed" if the operation or operations essential to recognition of that particular call were carried out suffi-

ciently well to receive credit. Thus, an Extension to Trunk call was considered to have been completed if the examiner had granted credit for the *ringing signal*, a Trunk to Busy Extension call if credit had been granted for the *busy report*, Transferring an Incoming Call if the *connection was maintained*, etc

The actual steps followed in scoring the test are enumerated in Exhibit C, which is a reproduction of the instructions furnished the scorers. Two copies of the scoring form (referred to in Exhibit C as EB-695) are reproduced as Exhibits D and E. The former is the record of an examinee who did not complete a sufficient number of calls to qualify in her county (Class II). The latter represents the rating of an examinee in a Class III County who completed considerably more than the minimum required in her county.

The use of the schedule of credits shown in Exhibit F made it possible to convert the approximately 75 sub-operations comprising the test into quantitative ratings.⁸ The maximum attainable score for the test designed for operators of dial equipment was 143, for operators of manual equipment, 123. To facilitate the scoring, keys were constructed which turned the task into a routine operation easily performed by clerks experienced in scoring objective tests.

Part II of this article will appear in the October issue of
EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

⁸The correlation between the number of calls completed and the score derived by applying the schedule of credits shown in Exhibit F is naturally quite high. In a test comprising 20 or 25 calls it would probably be unnecessary to go to the added trouble of weighting and scoring each part of a call. However, several considerations suggested the desirability of doing so in the particular test described.

PERFORMANCE TESTING IN PUBLIC PERSONNEL

Exhibit A
COMMONWEALTH OF PENNSYLVANIA
EMPLOYMENT BOARD
of the
DEPARTMENT OF PUBLIC ASSISTANCE
Harrisburg
PERFORMANCE TEST FOR TELEPHONE OPERATORS
SERIES 1000
August 1940

INSTRUCTIONS TO EXAMINEES

Important Failure to follow instructions may
result in disqualification from the examination

The examination you are about to take has been designed to test your ability to perform some of the tasks ordinarily required of a telephone operator in the Department of Public Assistance

When your turn arrives, you will be assigned to a PBX cord-type switchboard. The designation strips on this switchboard will indicate the location of four or more *trunk lines* and the following *extension stations*

Mr Albert
Mr Brown

Mr Carson
Mr Drake

Official

You will be given a few minutes to familiarize yourself with any aspects of the switchboard which are strange to you and to note the location of each of the jacks indicated above

When the Proctor tells you to do so, ring Mr Brown's telephone. Mr Brown will answer and ask you to repeat your Identification Number—the number which appears on your Admittance Slip

The examination, which consists of making various combinations of simple connections, will then begin. The first connection you will be required to make will be a practice exercise on which you will not be graded

When answering calls or acknowledging orders, the following phrases *must* be used.

Answering incoming calls—"Public Assistance "

Answering extension calls—"Yes, please?"

Acknowledging orders—"Thank you "

Note On incoming calls, if the Calling Party requests information regarding the Department of Public Assistance or asks to talk to anyone besides the four persons whose names are shown on the designation strip, the call must be connected to the extension marked "Official."

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Exhibit B COMMONWEALTH OF PENNSYLVANIA EMPLOYMENT BOARD of the DEPARTMENT OF PUBLIC ASSISTANCE Harrisburg

Identification No

Center
Date

Mr Albert ☐

Mr Brown ☐

Form D

PERFORMANCE TEST FOR TELEPHONE OPERATORS SERIES 1000

1 EXTENSION TO EXTENSION (Practice Exercise)

Mr Albert lifts receiver	
Operator answers	Promptness () "Yes, please?" ()
Mr Albert asks for Mr Brown "Mr Brown, please"	
Operator acknowledges	"Thank you" ()
Operator rings Mr Brown	Brown's phone rings ()
Mr Brown answers "Mr Brown speaking"	Connection completed ()
	Connection maintained ()
Mr Albert and Mr Brown hang up after a few seconds	

2. EXTENSION TO TRUNK

Mr Brown lifts receiver	
Operator answers	Promptness () "Yes, please?" ()
Mr Brown asks for city line "City line, please"	
Operator acknowledges	"Thank you" ()
Operator connects Mr Brown with trunk line	Dial tone () Promptness ()
Mr Brown dials listed number	Ringing signal ()
Mr Albert personally checks number of trunk line to which Mr Brown has been connected ()	

3 TRUNK TO EXTENSION WHICH DOES NOT ANSWER

Mr Brown's call comes in on trunk line	
Operator answers	Promptness () "Public Assistance" ()
Calling party (Mr Brown) asks for Mr Carson "Mr Carson, please"	
Operator acknowledges	"Thank you" ()
Operator rings Mr Carson	Carson's phone rings () Promptness ()
Operator gives ringing report Mr Brown tells Operator to continue ringing	Ringing report every 40 seconds () Appropriate phrase ()
	Connection maintained until transfer ()

PERFORMANCE TESTING IN PUBLIC PERSONNEL

Exhibit B (Continued)

4 EXTENSION TO BUSY EXTENSION, EXTENSION TO TRUNK

Mr Albert lifts receiver a few seconds after Mr Carson's telephone first rings	
Operator answers	Promptness () "Yes, please?" ()
Mr Albert asks for Mr Brown "Mr Brown, please"	
Operator gives busy report	Busy report () Promptness ()
Mr Albert asks for city line "City line, please"	
Operator acknowledges	"Thank you" ()
Operator connects Mr Albert with trunk line	Dial tone () Promptness ()
Mr Albert dials listed number	Ringing signal ()

5 TRUNK TO BUSY EXTENSION

Mr Albert's call comes in on trunk line	
Operator answers	Promptness () "Public Assistance" ()
Calling party (Mr Albert) asks for Mr Carson "Mr Carson, please"	
Operator gives busy report and asks calling party to hold line	Busy report () Hold line () Appropriate phrases ()
Calling party (Mr Albert) hangs up	

6 TRANSFERRING INCOMING CALL

Mr Albert lifts receiver	
Operator answers	Promptness () "Yes, please?" ()
Mr Albert asks to have Mr Carson's call "Let me have Mr Carson's call, please"	
Operator acknowledges	Appropriate phrase ()
Operator transfers incoming call from Mr Carson to Mr Albert	Appropriate phrase () Transfer () Promptness () Connection maintained ()
Mr Albert and Mr Brown hang up after a few seconds	

7 EXTENSION TO TRUNK

Mr Brown lifts receiver	
Operator answers	Promptness () "Yes, please?" ()
Mr Brown asks for city line "City line, please"	
Operator acknowledges	"Thank you" ()
Operator connects Mr Brown with trunk line	Dial tone () Promptness ()
Mr Brown dials listed number	Ringing signal ()
Mr Albert asks Operator number of trunk line to which Mr Brown has been connected	()

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Exhibit B (Continued)

8 TRUNK TO BUSY EXTENSION

Mr Brown's call comes in on trunk line	
Operator answers	Promptness () "Public Assistance" ()
Calling party (Mr Brown) asks for Mr Brown "Mr Brown, please "	
Operator gives busy report and asks calling party (Mr Brown) to hold line	Busy report () Hold line () Appropriate phrases ()
Calling party (Mr Brown) holds line	Connection maintained until transfer ()

9 TRANSFERRING OUTGOING CALL

Mr Albert lifts receiver	
Operator answers	Promptness () "Yes, please?" ()
Mr Albert asks to have call on Mr Brown's line transferred "Transfer the call on Mr Brown's line to me please "	
Operator acknowledges	Appropriate phrase ()
Operator transfers outgoing call from Mr Brown's telephone to Mr Albert's telephone	Appropriate phrase ()
Mr Albert listens for open line	Open line () Promptness ()
Mr Albert and Mr Brown hang up	

10 EXTENSION TO EXTENSION

Mr Brown lifts receiver	
Operator answers	Promptness () "Yes please?" ()
Mr Brown asks for Mr Carson "Mr Carson, please "	
Operator acknowledges	"Thank you " ()
Operator rings Mr Carson	Carson's phone rings ()
Mr Carson (Albert) answers "Mr Carson speaking "	Connection completed () Connection maintained ()
Mr Carson (Albert) and Mr Brown hang up after a few seconds	

11 EXTENSION TO TRUNK

Mr Albert lifts receiver	
Operator answers	Promptness () "Yes, please?" ()
Mr Albert asks for city line "City line, please "	
Operator acknowledges	"Thank you " ()
Operator connects Mr Albert with trunk line	Dial tone () Promptness ()
Mr Albert dials all but last digit of listed number and holds line	
Mr Brown asks Operator number of trunk line to which Mr Albert has been connected ()	

PERFORMANCE TESTING IN PUBLIC PERSONNEL

Exhibit B (Continued)

12 TRANSFERRING OUTGOING CALL

Mr Brown lifts receiver	
Operator answers	Promptness () "Yes, please?" ()
Mr Brown asks to have call on Mr Albert's line transferred	
"Transfer the call on Mr Albert's line to me, please "	
Operator acknowledges	Appropriate phrase ()
Operator transfers outgoing call from Mr Albert's telephone to Mr Brown's telephone	Appropriate phrase ()
Mr Brown listens for open line	Open line () Promptness ()
Mr Albert and Mr Brown hang up	

VOICE

To what extent is the Operator's voice clear, distinct, pleasant? 1 2 3 4
(1) Very unsatisfactory (2) Unsatisfactory, (3) Satisfactory, (4) Very satisfactory

REMARKS

(8-8 40)

Examiner

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Exhibit C

PROCEDURE FOR SCORING TELEPHONE OPERATOR PERFORMANCE TEST SERIES 1000

Note: All scoring must be checked and must carry the initials of both scorer and checker

- 1 Check to see that there are two rating sheets and an Admittance Slip for each examinee and that the Identification Number on each is identical
- 2 Write the examinee's Identification Number and County in the spaces provided on Form EB-695 (Use Form EB-696 for Form M)
- 3 Place a check mark on the rating sheet after the name of each call completed by the examinee
- 4 Place a check mark in the appropriate space on Form EB-695 for each completed call and enter the total number of calls completed in the box provided
- 5 Eliminate from further consideration examinees who completed fewer than the minimum number of calls required for their County (See attached schedule)
- 6 Score the rating sheets of examinees who completed a sufficient number of calls Place the number of credits after each line and place the total number of credits for each call in a circle to the right of the last line of the call (See attached schedule of credits.)
- 7 Transfer the number of credits for each call to the appropriate space on Form EB-695
- 8 Place a check mark or an "X" in each of the three spaces provided after the word "Trunks" on Form EB-695 and refer to the attached schedule for the number of credits to be entered in the space to the right
- 9 Place a check mark in the appropriate spaces after the word "Voice" on Form EB-695 and refer to the attached schedule for the number of credits to be entered in the space to the right
- 10 Enter the total number of credits earned (Raw Score) in the box provided on Form EB-695

PERFORMANCE TESTING IN PUBLIC PERSONNEL

Exhibit D

B-67432	Luzerne	File No.
Identification No.	Legal County	

(For use with Form D of test.)

Calls Completed	
Albert	Brown
4	X 2 ✓
4'	✓ 3 ✓
5	X 7 ✓
6	X 8 X
9	X 10 ✓
11	✓ 12 X

Trunks 1 2 3
Voice A3 A4 B3 B4

Raw Score **F**

EB-695

Exhibit E

B-20445	Erie	File No.
Identification No.	Legal County	

(For use with Form D of test.)

Calls Completed	
Albert	Brown
4	✓ 2 ✓
4'	✓ 3 ✓
5	✓ 7 ✓
6	X 8 ✓
9	✓ 10 ✓
11	✓ 12 ✓

Trunks 1 X 2 X 3 ✓
Voice A3 ✓ A4 B3 ✓ B4 4

Raw Score **104**

EB-695

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Exhibit F

PROCEDURE FOR SCORING TELEPHONE OPERATOR PERFORMANCE TEST
SERIES 1000

SCHEDULE OF CREDITS

Forms D and M*

CREDIT—1 POINT

Proper phrase
"Yes, please?"
"Public Assistance"
"Thank you"

Promptness
Appropriate phrase
Ringing report
Request to hold line

CREDIT—2 POINTS

Ringing extension telephone

CREDIT—4 POINTS

Busy report
(M) Central answers

CREDIT—6 POINTS

Extension-to-extension connection maintained
Note If connection is completed but not maintained, 3 points
(D) Ringing signal (after dialing)
(D) Open line (call 11 only)

CREDIT—8 POINTS

Incoming call transferred (connection maintained)
Note. If transfer is made without maintaining connection, 4 points
(M) Outgoing call transferred (open line)

CREDIT—10 POINTS

(D) Outgoing call transferred (open line—calls 9 and 12 only)

USE OF HIGHEST NUMBER TRUNK LINE

Once (D) 0 (M) 3
Twice (D) 5 (M) 8
Three times (D) 12

VOICE

Satisfactory—2 (each observer)
Very satisfactory—3 (each observer)

PASSING POINTS

Class II 10 completed calls
Class III 9 completed calls
Class IV 8 completed calls

*"D" in parenthesis indicates that the credit applies only to Form D,
"M" in parenthesis indicates that the credit applies only to Form M

SOME DATA ON THE KUDER PREFERENCE RECORD

ARTHUR E. TRAXLER
Educational Records Bureau
AND

WILLIAM C. MC CALL
University of South Carolina

A WELL-ROUNDED guidance program calls for at least four types of objective measures—general intelligence, achievement in various fields of study, aptitudes of different types, and interests or motivation. Far more progress has been made in the first three of these areas than in the fourth. In recent years, however, there has been an especially large amount of experimentation in the last area and some promising measuring instruments are beginning to emerge.

The majority of the noteworthy instruments for appraising interests have been concerned with occupational preferences. The most important work in this field has been done by Strong, who has constructed blanks and prepared scales for the measurement of the interests of men with respect to 34 occupations and the interests of women in connection with 18 occupations. Although the instruments developed for the measurement of interest in specific vocations unquestionably have important guidance values, at least two considerations point to a trend away from the measurement of interests in occupations as such and toward the measurement of interests in broad fields.

One consideration is based on observation and research. It has been known almost from the first attempts to measure vocational interests that interests in certain vocations are rather highly correlated. It has been appar-

ent that there are clusters of occupations that have so many points of similarity that interest in one occupation is a strong indication of interest in several others. Factor-analysis studies have given emphasis to this point. For example, by means of a factorial analysis of the Strong Vocational Interest Blank, Thurstone¹ found four interest groups. These groups were associated with science, language, people, and business.

The second consideration grows out of the practical, everyday work of counselors and personnel officers. These workers have found that frequently when one is attempting to guide the development of secondary-school pupils, or even of college freshmen, guidance with respect to specific occupations is not needed. In fact, guidance into specialization so early would in many cases be unwarranted. What is needed is a valid, reliable measure of interests in fairly broad fields so that the individual may be guided in the general direction of a group of related occupations, one of which will perhaps be chosen definitely when the student has attained greater maturity.

Strong, himself, has been one of the first to recognize the need for broader measurement of interests as well as measurement of interests related to specific vocations. In line with this viewpoint, he has recently published several group scales for the measurement of interests in broad areas.

Certain other investigators have been working along somewhat similar lines. Probably the most promising new instrument in this general field is the *Preference Record* by G. F. Kuder.²

Description of the Preference Record

The Preference Record is designed for use in obtaining measures of motivation in the following seven fields: scientific, computational, musical, artistic, literary, social

¹L. L. Thurstone, "A Multiple Factor Study of Vocational Interests," *Personnel Journal*, X (1931), 198-205.

²G. Frederic Kuder, *Preference Record* (Chicago: Science Research Associates, 1939).

DATA ON KUDER PREFERENCE RECORD

service, and persuasive. It consists of 330 paired-comparison items of which the following are samples

- A (1) Draw graphs
- (2) Do clerical work
- B (1) Be a lawyer
- (2) Be a landscape architect
- C (1) Sell insurance
- (2) Do scientific research work

The subject indicates in each case which one of the pair of activities he prefers.

The test is intended for use in high school and college. It is administered without time limit. The booklet is used with separate answer sheets, one for hand scoring, one for machine scoring, and one for self scoring. The raw scores of an individual student may be plotted on a percentile chart and thus a graphic indication of high points and low points with respect to the seven fields may be obtained.

Nature and Purpose of the Study

Kuder³ has described the construction of the Preference Record in some detail and has reported a considerable amount of statistical data for it. Helpful as these data are, they naturally do not cover all questions about the blank. Since no other studies of this new instrument were available, it seemed desirable to try to obtain answers to certain questions before arriving at decisions about the use of the blank in a regular testing program. The questions which this study attempts to answer are as follows.

- 1 What is the retest reliability of the scores on the Kuder Preference Record?
- 2 Are the scores on the Preference Record relatively stable over a long period?
- 3 What differences are there between the mean scores for boys and for girls on the Preference Record?

³G. F. Kuder, "The Stability of Preference Items," *The Journal of Social Psychology*, X (1939), 41-50.

- 4 Do the mean scores for different secondary-school groups change appreciably with change in grade level?
- 5 What is the shape of the mean profiles of university freshmen in different fields of study?

The data were obtained by administering the Preference Record to freshmen in the University of South Carolina, pupils in Grades 10, 11, and 12 of a high school in South Carolina, and a number of adults who were on the staff of an educational organization in New York City

Reliability

In the manual for the Preference Record, Kuder gives the following reliabilities for the different scales: scientific, .87; computational, .85; musical, .88; artistic, .90; literary, .90; social service, .84; persuasive, .90. These reliabilities were estimated from one administration of the test to a group of 84 college students through the application of the Kuder-Richardson method of estimating reliability coefficients.⁴ Since the procedure employed is still somewhat experimental and is not as yet generally used, it seemed advisable to check the reliability of the various scales by the more familiar test-retest procedure. Accordingly, 52 college freshmen and 90 high-school pupils who had filled out the Preference Record near the beginning of the term were retested after an interval of a few weeks. The elapsed time was approximately one month for the high-school pupils and two months for the college students. The correlations between the scores resulting from the two administrations are shown in Table 1. Means and standard deviations of the distributions are also given.

For all scales, the correlations between the two administrations of the Preference Record to the secondary-school group are above .8. They vary from approximately .81 to about .91. With the exception of the correlation

⁴G. F. Kuder and M. W. Richardson, "The Theory of the Estimation of Test Reliability," *Psychometrika*, II (1937), 151-60.

DATA ON KUDER PREFERENCE RECORD

TABLE 1
RETEST RELIABILITY OF THE KUDER PREFERENCE RECORD BASED
ON THE SCORES OF SECONDARY SCHOOL PUPILS AND OF COLLEGE
FRESHMEN IN SOUTH CAROLINA

	Secondary School Pupils							College Freshmen						
	N	r	PE	Mx	SDx	My	SDy	N	r	PE	Mx	SDx	My	SDy
Scientific	90	907±	013	41 80	9 33	41 83	9 25	52	782±	036	42 08	9 68	43 65	9 19
Computa- tional	90	814±	024	19 82	7 24	19 38	6 75	52	748±	041	18 15	7 88	18 73	6 70
Musical	90	876±	017	16 69	7 39	16 24	7 18	52	871±	023	18 92	7 15	17 54	6 49
Artistic	90	857±	019	30 87	8 97	31 10	8 04	52	820±	031	30 92	8 86	29 85	8 36
Literary	90	863±	018	32 30	9 71	32 80	10 14	52	789±	035	31 56	10 10	34 27	10 67
Social														
Service	90	838±	021	39 97	9 66	41 20	10 15	52	588±	061	41 37	9 42	43 54	8 02
Persuasive	90	838±	021	15 27	9 25	46 27	8 95	52	795±	034	45 62	9 30	46 65	9 16

for the social service scale, the correlations between the two administrations of the test to the college group are above .74. They range upward to approximately .87. In general, these coefficients are rather high for correlations based on retesting after an interval of several weeks. In fact, most of the correlations seem exceptionally satisfactory for a measuring device that can be administered and scored so quickly and that yields as many as seven scores.

The correlations based on the secondary-school group are high enough to warrant considerable use of the Preference Record in individual prediction and guidance. Reliability coefficients above .90 are theoretically desirable for a test that is to be used in this way, but experience indicates that they are very seldom attained in a test that yields several different scores.

The correlations for the college freshmen tend to be lower than those for the secondary-school pupils, although there is no significant difference in the case of the musical scale. The correlation for the social service scale, .588, is much the lowest in the group. The secondary-school data indicate, however, that this scale is not less reliable than some of the others.⁵

The reason for the somewhat higher correlations at the secondary-school level than at the college-freshman level is not entirely clear. It was thought at first that pos-

⁵Since this correlation was out of line with the others, it was rechecked with the greatest of care. Every paper was rescored, the data were redistributed, and the entire calculation was carried through a second time. It seems certain, therefore, that no error of a clerical nature is involved.

sibly the combining of the three high-school grades into one group had increased the variability and thus raised the correlations. However, a comparison of the standard deviations shows that in general the variability is not greater for the high-school group.

The difference in magnitude of the correlations may be due simply to the longer time interval between administrations of the Preference Record to the college group. By the same reasoning, the correlations for the high-school pupils may be a little lower than they would have been if only a few days had elapsed before the test was repeated. It is improbable, however, that the basic interests and motives of either high-school or college students change significantly during a period of a few weeks. Moreover, the repetition of the Preference Record after a very brief period would have been subject to the limitation that a memory factor might have produced spuriously high correlations.

The retest correlations based on the secondary-school group correspond rather well with the estimated reliabilities reported by Kuder. The correlation obtained in this study for the scientific scale is a little higher than Kuder's figure. The two sets of reliabilities for the musical scale and the social service scale would agree exactly if those reported here were rounded to two decimal places. In the case of the other four scales, the retest correlations for the secondary-school pupils are lower than Kuder's reliabilities, but the differences are not marked.

The college-freshman retest correlation for the musical scale is in very close agreement with Kuder's reliability coefficient. The college-freshman correlations for the other scales are significantly lower than those found by Kuder, but the only striking difference is between the correlations for the social service scale.

Means and Standard Deviations

Although this is not concerned with one of the main questions raised in this study, it may be noted in passing

DATA ON KUDER PREFERENCE RECORD

that the means and the variabilities of the distributions resulting from the two administrations of the Preference Record tend to be closely similar in both groups. Apparently the practice effect was negligible, that is, the scores did not tend to be higher on the second administration as a result of the subjects' having taken the test previously. The absence of evidence of practice effect is a further point in favor of the Preference Record.

Another interesting observation based on the means and standard deviations shown in Table 1 is that, on the whole, the difference between the two groups in central tendency and variability are slight. This observation suggests that interests in the seven areas involved are relatively mature by the time pupils enter the secondary school. The largest difference in favor of the college-freshman group is found in the social service scale, a result which familiarity with scores of high-school and college students on the Strong Vocational Interest Blank would lead one to expect.

Stability of the Scores

We have just noted that the retest correlations for the scales of the Kuder Preference Record tend to be rather high for an interval of a few weeks. But how high would they be for a rather long period—let us say, a year or more? In other words, what is the stability of the scores and what is their value for long-time predictions? Some information relative to these questions is provided by the correlations in Table 2, which are based on the retesting

TABLE 2
CORRELATIONS BETWEEN SCORES MADE BY SIXTEEN ADULTS ON TWO ADMINISTRATIONS OF THE KUDER PREFERENCE RECORD AFTER AN INTERVAL OF APPROXIMATELY FIFTEEN MONTHS

Scale	N	r PE	Mx	SDx	My	SDy
Scientific	16	.828±.053	47.25	12.14	47.50	10.94
Computational	16	.864±.043	24.56	8.27	23.81	8.68
Musical	16	.933±.022	22.00	8.46	22.63	8.31
Artistic	16	.698±.086	31.38	6.90	31.25	8.35
Literary	16	.810±.058	44.25	8.48	44.25	8.18
Social Service	16	.611±.106	42.38	7.17	41.50	8.29
Persuasive	16	.883±.037	31.38	11.70	32.88	12.42

of 16 adults with the Preference Record after an interval of approximately 15 months

The correlations in Table 2 range from above .6 for the social service scale to above .9 for the musical scale. The correlations for all the scales except the artistic and social service ones are above .8. The reliability of these correlations is of course limited by the small number of cases. For example, the correlation coefficient for the artistic scale was lowered considerably by a marked change in the score of one person.

Nevertheless, the correlations in Table 2 suggest that in general the scores on the Preference Record are rather stable for a period as long as 15 months and that they provide a fairly satisfactory basis for long-time predictions. Emphasis is given to this point when one examines the preference profiles of the different individuals. In nearly all cases, the high and the low points resulting from the first administration of the test were closely similar to those based on the second administration. The profiles for two individuals, in terms of percentiles, are shown in Figures 1 and 2.

Sex Differences

When one is interpreting the profile of an individual on the Preference Record, it is of some interest to know whether there are characteristic differences between the average scores of boys and girls. The mean scores made on the various scales of the Preference Record by groups

TABLE 3
MEAN SCORES OF GROUPS OF BOYS AND GIRLS IN HIGH SCHOOL
AND IN COLLEGE ON THE KUDER PREFERENCE RECORD

Scale	High School Boys	High School Girls	Freshman Boys in a State University	Freshman Girls in a State University	Freshmen in a Girls' College
Number of Cases	152	135	303	173	584
Scientific	48.1	37.9	47.9	41.4	41.7
Computational	21.6	16.9	21.8	17.7	17.8
Musical	13.2	18.2	15.8	19.7	20.6
Artistic	27.4	29.9	26.9	32.1	29.7
Literary	30.4	33.4	32.8	36.7	35.8
Social Service	37.8	44.8	40.7	45.4	46.7
Persuasive	47.9	44.9	49.4	42.8	44.0

DATA ON KUDER PREFERENCE RECORD

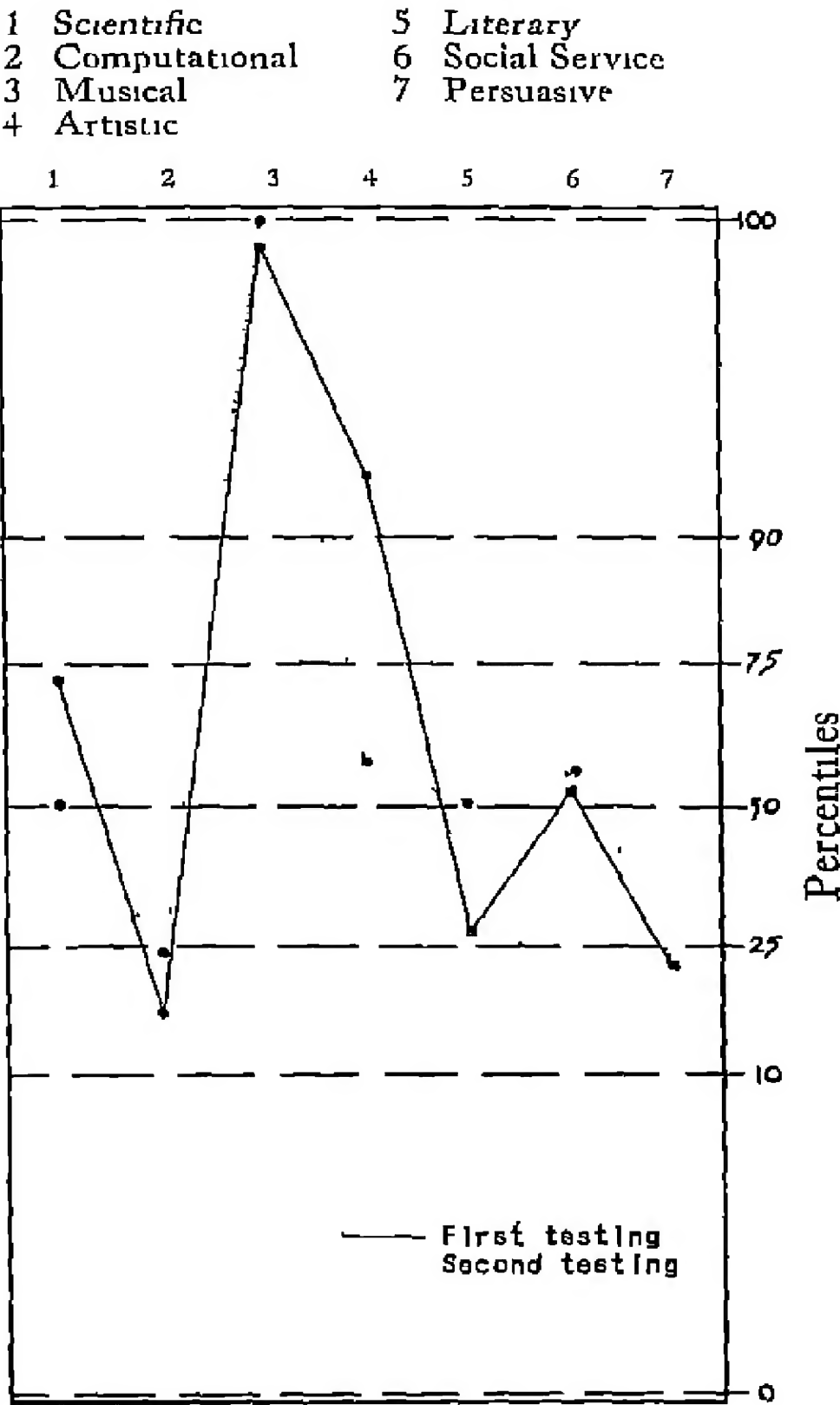


Figure 1 Profile of a Girl Secretary with a Long-Standing Interest in Music

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

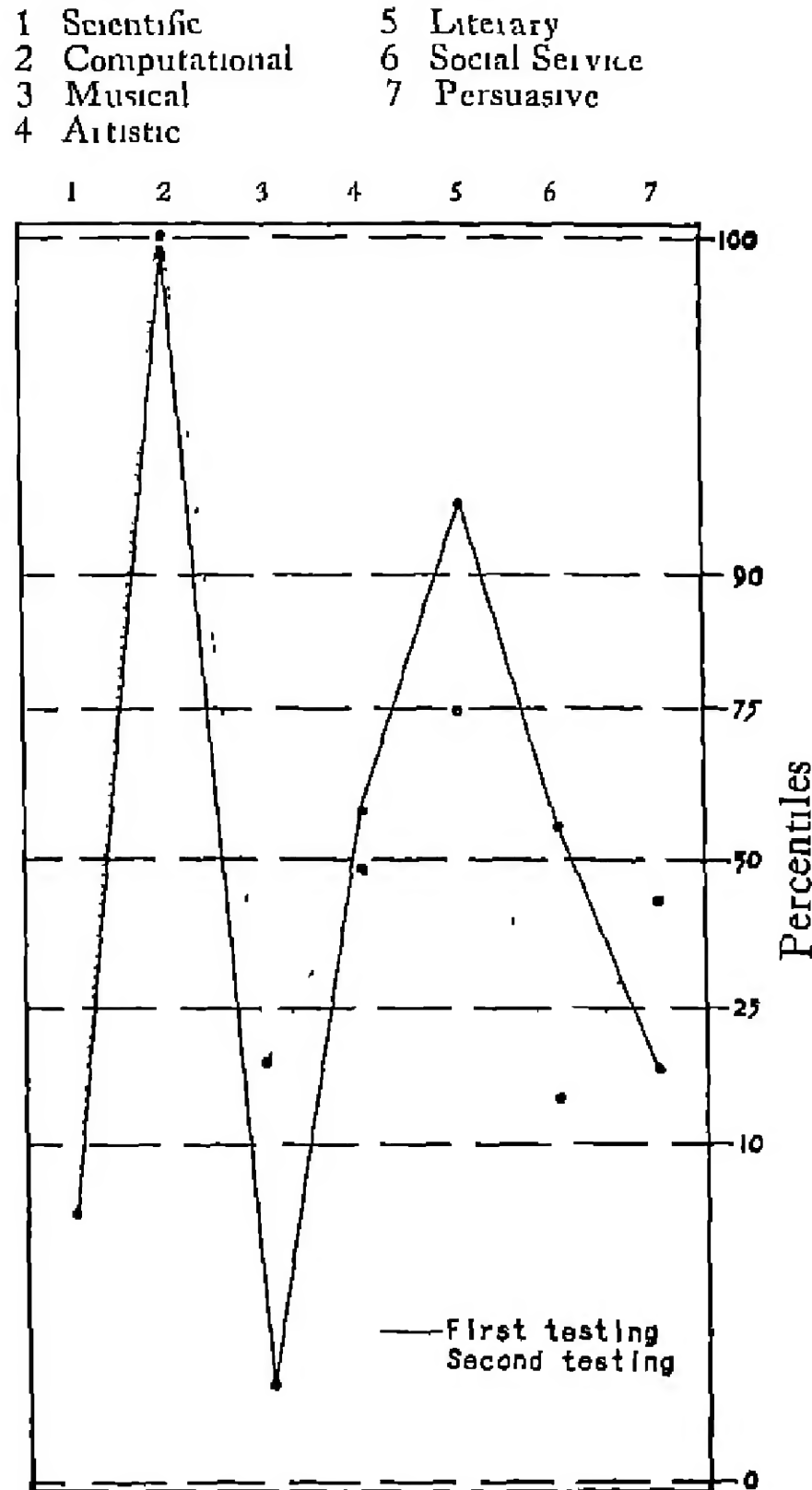


Figure 2 Profile of a Machine Scoring Supervisor, an Occupation for which Interest and Ability in Computation Are Very Important

DATA ON KUDER PREFERENCE RECORD

of high-school boys and girls and college-freshman boys and girls in South Carolina are presented in Table 3

As one might expect, in both groups the boys are on the average higher than the girls in scientific, computational, and persuasive preferences, while the girls surpass the boys in musical, artistic, literary, and social service preferences. The largest differences are for the scientific scale. However, even the smallest differences in medians amount to nearly 10 percentile points. It appears, therefore, that sex differences in preferences should be taken into consideration when the scores on this test are interpreted.

Grade-Level Differences

When the results of achievement tests are being studied, the usual procedure is to interpret the scores in terms of the norms for the grade the pupils are in. Is it necessary to follow this procedure with the Preference Record or are the results in different grades so similar that one is justified in disregarding grade level, at least as far as the secondary school is concerned? Some information on this question is given in Table 4, which shows mean scores of boys and of girls in Grades IX, X, and XI of a South Carolina High School.

TABLE 4
MEAN SCORES MADE ON THE PREFERENCE RECORD BY GROUPS OF BOYS AND GIRLS IN GRADES IX, X, AND XI OF A SOUTH CAROLINA HIGH SCHOOL

Scale	Boys			Girls		
	Grade IX	Grade X	Grade XI	Grade IX	Grade X	Grade XI
Number of Cases	48	77	27	42	72	21
Scientific	48.8	48.4	45.8	36.1	39.0	38.0
Computational	20.5	22.0	22.2	16.9	16.9	16.7
Musical	11.3	13.6	15.2	18.9	17.4	19.8
Artistic	27.1	28.1	25.9	31.3	29.4	28.8
Literary	30.3	30.5	30.1	32.9	33.9	32.9
Social Service	40.0	37.0	36.3	44.0	44.6	47.1
Persuasive	47.8	47.5	49.2	42.3	46.0	46.4

Because of the rather small number of cases, the means are not highly reliable indicators of the preferences at the different grade levels. Nevertheless, the fluctuations in

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

mean scores made by the pupils in the three grades are not great. Moreover, there is no consistent trend toward the obtaining of higher or lower ratings with advancement in grade level. The data in Table 4 are by no means conclusive, but, as far as they may be interpreted, they suggest that different norms for these three grades are not needed.

Mean Profiles for Different Fields of Study

The means and standard deviations of scores of University of South Carolina freshmen classified according to field of specialization are shown in Table 5. The percentile ratings of the mean scores are indicated graphically in Figure 3.

TABLE 5

MEANS AND STANDARD DEVIATIONS OF SCORES MADE ON KUDER PREFERENCE RECORD BY VARIOUS GROUPS OF FRESHMEN IN THE UNIVERSITY OF SOUTH CAROLINA, INCLUDING BOTH MEN AND WOMEN

Scale	Engineering (B S)			Journalism			Art			Education (A B)		
	N	Mean	S D	N	Mean	S D	N	Mean	S D	N	Mean	S D
Scientific	79	53.30	6.20	26	39.23	6.10	15	41.93	7.19	36	39.50	10.08
Computational	79	34.29	5.32	26	11.92	5.48	15	17.67	10.52	36	16.83	6.66
Musical	79	14.44	6.45	26	18.77	6.11	15	18.07	5.31	36	18.72	7.85
Artistic	79	29.18	7.17	26	25.23	7.22	15	45.67	7.33	36	25.50	7.98
Literary	79	29.38	7.44	26	54.16	6.55	15	31.13	5.54	36	38.28	8.69
Social Service	79	40.14	7.21	26	41.00	9.38	15	44.87	5.08	36	44.78	11.46
Persuasive	79	46.90	8.06	26	18.15	9.33	15	42.73	8.79	36	44.28	9.48

Scale	Commerce (B S)			Secretarial			Science			Pre Medicine			Pharmacy		
	N	Mean	S D	N	Mean	S D	N	Mean	S D	N	Mean	S D	N	Mean	S D
Scientific	82	10.85	8.50	83	38.59	8.03	53	56.24	8.66	13	54.69	7.80			
Computational	82	21.71	6.47	83	20.86	7.11	53	17.04	5.31	13	16.54	5.21			
Musical	82	15.19	6.47	83	19.12	5.89	53	16.81	7.22	13	18.54	8.16			
Artistic	82	24.98	6.16	83	30.18	8.11	53	26.77	7.02	13	28.08	6.82			
Literary	82	31.12	8.53	83	33.84	8.55	53	34.28	7.95	13	30.08	7.13			
Social Service	82	40.85	7.79	83	43.00	8.99	53	48.59	7.25	13	42.85	7.08			
Persuasive	82	53.83	8.66	83	45.17	9.16	53	44.62	10.55	13	43.62	8.39			

Scale	Pre Law			Arts and Sciences (A B)			Arts and Sciences (B S)		
	N	Mean	S D	N	Mean	S D	N	Mean	S D
Scientific	32	40.00	3.50	169	41.43	8.60	49	51.45	9.41
Computational	32	20.69	5.85	169	17.64	6.68	49	19.57	6.98
Musical	32	16.69	5.22	169	19.13	7.32	49	16.18	7.17
Artistic	32	24.13	7.38	169	30.66	9.30	49	27.61	8.32
Literary	32	38.00	8.65	170	35.71	10.21	49	34.43	8.59
Social Service	32	38.13	6.34	169	44.33	10.38	49	43.12	10.07
Persuasive	32	57.94	10.99	170	44.29	10.42	49	44.10	10.65

The profiles of no two groups are alike. The greatest similarity probably is in the profiles for the pre-medical and pharmacy groups, but even here the correspondence is not especially close when all the scores are considered. Most of the high points and low points in the profiles

DATA ON KUDER PREFERENCE RECORD

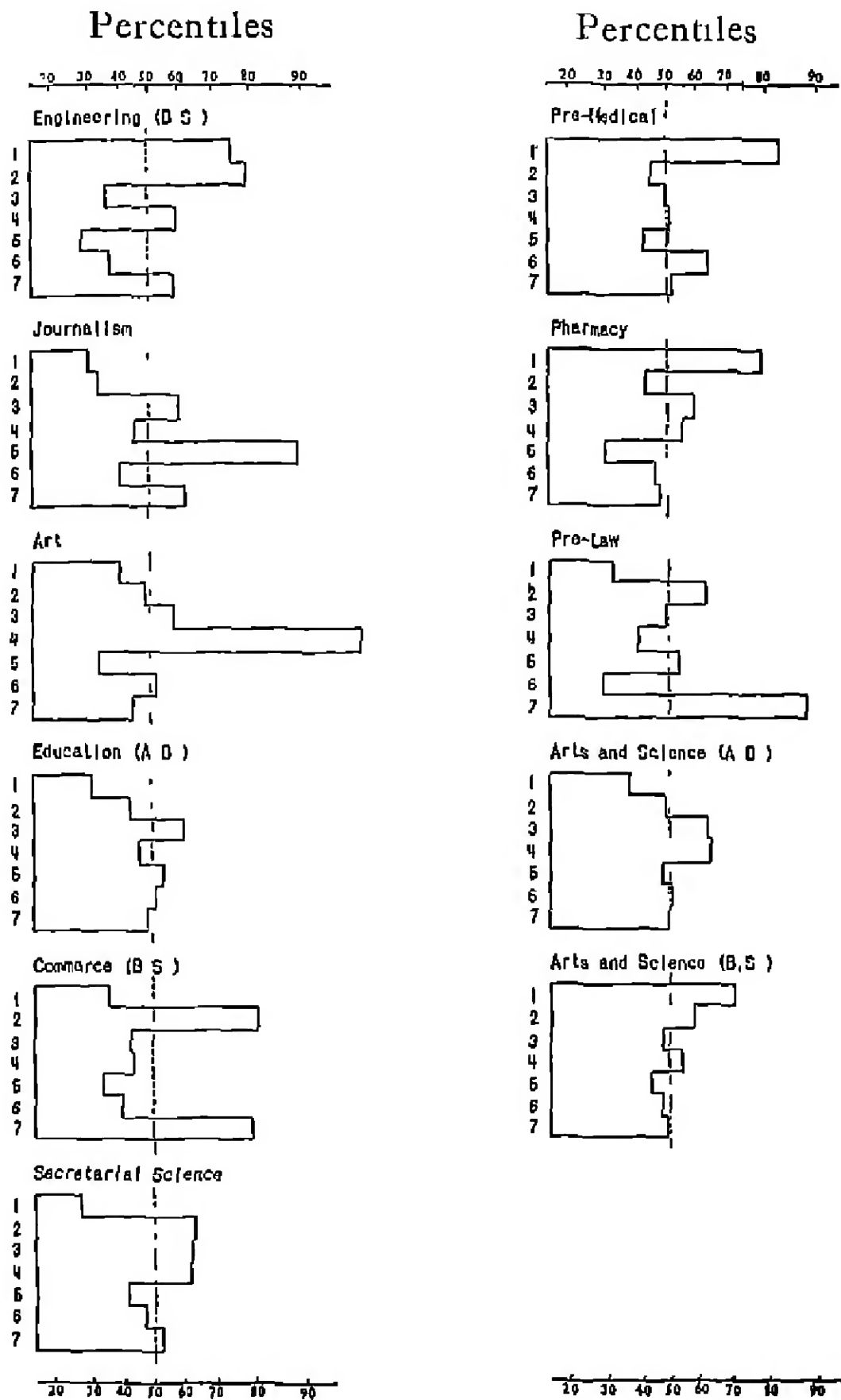


Figure 3 Mean Profiles of Groups of Freshmen Who Have Indicated Educational or Occupational Choices in the Fields Named

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

occur according to one's expectation. The art group, for example, is rather low in the scientific scale but very high in the artistic scale. The journalism group is close to the thirtieth percentile on the scientific and computational scales, but almost up to the ninetieth percentile on the literary scale. The commerce group is approximately at the eightieth percentile in computational and persuasive interests, but close to the thirty-fifth percentile in scientific and literary interests. Both the pre-medical and the pharmacy students are high in scientific interests. The pre-medical students are also fairly high in social service interests. The pre-law students are close to the thirtieth percentile on the scientific and social service scales, but near the eighty-fifth percentile on the persuasive scale and not far below the seventieth percentile on the computational scale. The engineering group is above the seventieth percentile in computational and scientific interests, but below the thirtieth percentile in literary interests.

In the manual for the Preference Record, Kuder has given median profiles for groups of students who have chosen occupations in the fields of writing, social service, physical sciences, political science, business and accounting, veterinary medicine, medicine, and law. A comparison of the profiles in Figure 3 with Kuder's profiles reveals noteworthy similarities between those for (1) journalism and writing, (2) commerce and business and accounting, (3) pre-medical course and medicine, (4) pre-law and law, and (5) arts and sciences (B S) and physical sciences. The fact that the profiles derived from two independent sources for groups in the same general areas are similar and are on the whole in agreement with what one would reasonably expect is favorable to the reliability and validity of the Preference Record.

Conclusions

1. The retest reliability of the scales of the Kuder Preference Record is rather high. The correlations between the scores resulting from two administrations of the

DATA ON KUDER PREFERENCE RECORD

Preference Record to a group of high-school pupils with a time interval of about one month were above .8 for all seven scales. The correlations between the scores based on two administrations of the Record to a group of college freshmen with a time interval of two months were above .7 for six of the seven scales.

2. The scores on the Preference Record do not seem to be influenced by practice in taking the Record when there is an interval of several weeks between administrations of the record. The mean scores resulting from the second administration of the Record were not appreciably or consistently higher than the scores obtained the first time the Record was taken.

3. The scores on the Preference Record appear to have considerable value for relatively long-time predictions as far as adults are concerned. The correlations between the scores of 16 adults after an interval of 15 months were fairly high, varying from about .6 to slightly above .9.

4. There are noteworthy sex differences between the mean scores of high-school and college boys and girls. On the average, the boys exceed the girls in scientific, computational, and persuasive preferences; the girls are higher than the boys in musical, artistic, literary, and social service preferences.

5. It appears that interests and motivation in the seven areas involved are relatively mature by the time pupils reach the secondary school. The differences between the mean scores of pupils in Grades IX, X, and XI were found to be slight, and there was no consistent trend toward higher scores with increase in grade level. Similarly, the differences between the mean scores of the secondary-school and college freshman groups were small.

6. Mean profiles were found for 11 groups of university freshmen classified according to field of study or occupational choice. The profiles tended to have high points and low points at the places where one would

expect them to be. A comparison with profiles found by Kuder showed that those for groups in the same general fields were similar in shape.

In general, the data in this article are favorable to the Kuder Preference Record. By far the most important question that remains to be answered has to do with the validity of the Record. Do the scales really measure what they purport to measure? While certain aspects of the data reported in Kuder's manual and in this article imply considerable validity, there is at present little direct evidence concerning the validity of the Preference Record. One of the writers reported a small amount of data on the validity of the Record in Buross' *1940 Mental Measurements Yearbook*, but there was nothing conclusive in the findings. Further study of the Preference Record could well be directed toward this question.

THE RELIABILITY OF RATIO SCORES

LEE J. CRONBACH
State College of Washington

EDUCATIONAL measurements often give rise to quotients or ratios obtained when one score is divided by another. The intelligence quotient, achievement quotient, and per cent accuracy scores are examples. For the effective interpretation of such a measure, it is important that an appropriate estimate of its reliability be obtained. While a formula for the reliability of ratios has been presented by Holzinger, this, like other approaches, has limitations which apparently have not previously been discussed. The present article is intended to summarize the procedures which may be applied to ratio scores and to indicate the conditions under which each is appropriate.

Ratio scores appear to be particularly important in dealing with certain new-type tests such as those now being published by the Progressive Education Association. In Test 141, *Social Problems*,¹ for example, the student is presented with a description of a social problem, asked to state which of several solutions he favors, and then to check, from a long list, which reasons he would advance to support his decision. Since the student may check as many reasons as he wishes, there is in practice a wide range of "Total Reasons" among students. One important datum is the extent to which the student checks reasons which are inconsistent with his conclusion.

¹*Test 141, Social Problems*. Chicago: Progressive Education Association.

and really support one of the other solutions. This is expressed in the score "Number Inconsistent." In order to compare one student with another, it is convenient to eliminate the comprehensiveness factor, expressing his performance in a "Per Cent Inconsistent" score by dividing Number Inconsistent by Total Reasons.

Retest method One of the most satisfactory estimates of the reliability of such a score is to be obtained by the retest method. This method has generally been used to determine the reliability of the I-Q. It is not easy to rule out possible practice effect, even when parallel forms are used. For some tests it is difficult to prepare a parallel form, even where two forms are available, it is often desired to estimate the reliability without the trouble a second testing requires. Once data from two forms are available, it is a simple matter to compute the two ratios for each student and correlate them. It will be shown below, however, that a coefficient so obtained may not be equally appropriate for all scores in a given population.

Kuder-Richardson method Where retesting is impracticable, it is customary with ordinary tests to use the Spearman-Brown split-half procedure or the recently developed Kuder-Richardson method. The Kuder-Richardson method is based upon a summation of the variance of the items composing the total score,² since a ratio score cannot be conceived as composed of a sum of items, the method is not applicable (except in that special case where the denominator of the ratio is a constant for all students). Whether a modification of this method can be developed which is appropriate for ratios is not known.

Split-half method The Spearman-Brown formula is based on the assumption that the two halves into which the test has been split may be added to form the whole. In the case of ratio scores, this assumption does not hold. Since the denominator of the ratio is, in general, different

²G. F. Kuder and M. W. Richardson, "The Theory of the Estimation of Test Reliability," *Psychometrika*, II (1937), 151-60.

RELIABILITY OF RATIO SCORES

in each half, $\frac{a_{1/2}}{b_{1/2}} + \frac{a_{2/2}}{b_{2/2}}$ is not equal to $\frac{a}{b}$, or $1/2 \frac{a}{b}$. It would be possible to correlate $\frac{a_{1/2}}{b}$ with $\frac{a_{2/2}}{b}$, this would, by the Spearman - Brown formula, yield $\frac{1}{2} \frac{a_1 a_2}{b b}$. Since this disregards the possibility of error in measuring the denominator, it does not estimate the reliability of the ratio in the usual sense of $\frac{1}{2} \frac{a_1 a_2}{b_1 b_2}$. It is obvious that these objections do not apply to the special case where the denominator is the same for every student, as in the per cent accuracy score on a test where every student attempts every item. Here, the reliability of the ratio is the same as the reliability of the numerator.

Computation by formula. Statistical formulas for obtaining the mean, standard deviation, and correlations involving ratios by indirect methods were developed by early workers. These formulas, obtained by assuming that the variation of the denominator is small compared to its mean, are as follows

If $i = \frac{a}{b}$, $j = \frac{c}{d}$, $v_a = \frac{\sigma_a}{M_a}$, and so on, then

$$M_i = \frac{M_a}{M_b} (1 - r_{ab} v_a v_b + v_b^2), \tag{1}^3$$

$$\sigma_i^2 = \frac{M_a^2}{M_b^2} (v_a^2 - 2 r_{ab} v_a v_b + v_b^2); \tag{2}^3$$

$$r_{ij} = r_{\frac{a}{b} \frac{c}{d}} = \frac{r_{ac} v_a v_c - r_{ad} v_a v_d - r_{bc} v_b v_c + r_{bd} v_b v_d}{\sqrt{v_a^2 + v_b^2 - 2 r_{ab} v_a v_b} \sqrt{v_c^2 + v_d^2 - 2 r_{cd} v_c v_d}} \tag{3}^4$$

If the reliability of a score is conceived as its correlation with itself, one may substitute a for c and b for d in

³G. U. Yule and M. G. Kendall, *Introduction to the Theory of Statistics* (12th ed., revised, Philadelphia: Lippincott, 1937), pp. 299-300.

⁴Karl Pearson, "On a Form of Spurious Correlations Which May Arise When Indices Are Used in the Measure of Organs," *Proceedings of the Royal Society*, LX (1897), pp. 489 ff.

formula (3), obtaining this formula for the reliability

$$r_{aa} = \frac{r_{aa}v_a^2 - 2r_{ab}v_av_b + r_{bb}v_b^2}{v_a^2 - 2r_{ab}v_av_b + v_b^2} \quad (4)^5$$

This formula may be employed wherever the necessary data can be computed. Since r_{aa} , r_{bb} , and the other variables can be obtained without a retest, data from a single testing are sufficient. Either the split-half or Kuder-Richardson method may be used to estimate these reliabilities. It must be emphasized, however, that the formula is applicable only when the basic assumption is valid, namely, when the spread of the denominator variable is small compared to its mean.

In actual school testing of a single grade, the variation in mental age or chronological age is normally quite a small fraction of the mean M.A. or C.A. The ratio $\frac{\sigma}{M}$ for M.A. and C.A. is likely to be sufficiently small that higher powers can be neglected, as a result, the formula can be applied to either the A.Q. or the I.Q. under this condition. An empirical test of the formula by Morley showed close correspondence between formula results and results from a retest of 381 pupils, all in Grade VIIA, on several achievement quotients.⁶ It does not follow that the formula can be applied to the achievement quotient if several grades are included in one population.

When the Per Cent Inconsistent score is studied, one finds that the assumption does not necessarily hold. The average student checks less than 50 reasons, but the range in Total Reasons often is from 10 reasons to 70 reasons. The coefficient of variation of the denominator in such a case is so high that error may follow when the formula is applied. A further confusion lies in the fact that all

⁵This formula was first developed by Holzinger. See K. J. Holzinger, "Formulas for the Correlation between Ratios," *Journal of Educational Psychology*, XIV (1923), 344-47. It may also be derived directly by approximation from expansions of infinite series.

⁶C. A. Morley, "The Reliability of the Achievement Quotient," *Journal of Educational Psychology*, XXI (1930), 355-56.

RELIABILITY OF RATIO SCORES

scores in a given population are not equally reliable. A digression to demonstrate this point is necessary before methods of attacking this problem can be presented.

It is well known that the reliability of a test is a function of the length of the test. The student who marks three inconsistent reasons out of 10 reasons used, and the student who marks 30 inconsistent reasons out of the 100 he uses, both receive Per Cent Inconsistent scores of 30. The score of the latter student is an estimate of his inconsistency based on 100 responses, the estimate of the former student is based on only 10. If the former student were to mark one additional reason, his Per Cent Inconsistency score would increase to 36 per cent or decrease to 27 per cent, if the second student were to mark one more reason, his score would shift upward only to 30.7 per cent or downward to 29.7 per cent, depending, of course, on whether the additional reason were consistent or not. Similarly, a change of only one point in the numerator produces a much greater change in the ratio for the student whose denominator score is low. From a logical point of view, then, we would expect the standard error of measurement of a ratio score to increase as the denominator decreases. The standard error of measurement and the reliability coefficient are inversely related; therefore, the reliability of a score increases as the size of the denominator increases.

Possibly reference to the per cent accuracy concept will further clarify this point. If we were to ask a student a single question, he could answer it correctly or incorrectly or could omit it. If we desired to know the percentage of his attempts that were successful, we could compute a per cent accuracy score, which, based on one question, could only be 100, 0, or indeterminate. Certainly no measure of this sort, based on one item, would be considered significant. If two questions were asked, he could have both right, both wrong, one right and one wrong, or could omit either, or both. His per cent accuracy score,

under each of these conditions in order, would be 100, 0, 50, 100 or 0, or indeterminate. Obviously, when a score of 50 per cent is possible, discrimination is finer than when only scores of 100 and 0 are possible. Similarly, as the number of items *attempted* increases, discrimination becomes increasingly fine, which means that accuracy, hence reliability, of measurement increases. No matter how many items are added to the test, if a student omits all the items no meaningful per cent accuracy score can be obtained, and, in general, the accuracy with which his performance is measured depends upon the number of items he attempts. Timble has suggested⁷ that this applies not only to the ratio scores, but to scores on any test where the student is instructed to respond to as many items as he wishes, this pattern is found in several tests of the Progressive Education Association series.

Since it has therefore been demonstrated that the reliability of any score $\frac{a}{b}$ is a function of the size of b , as well as of the test used and the group measured, one may raise the question: how can a formula for reliability of a ratio, giving a single answer, be meaningful? The answer may be obtained by recalling the basic assumption under which the formula was obtained, to wit, that it holds only for those cases where σ_b is small compared to b . This is of course most likely where either (a) there is little variation in b scores within the group or (b) values of b are high. In the former case, all scores will have about the same reliability, which can be estimated by formula (4). In the latter case, the value obtained by the formula is a limiting value. It was pointed out that the reliability of a ratio increases as the denominator increases, other things being equal. Since, as the denominator increases, the case is approached where powers of $\frac{\sigma_b}{b}$ are completely negligible,

it follows that the value given by formula (4) is valid

⁷In correspondence with the writer

RELIABILITY OF RATIO SCORES

only where the assumption is met, and that for lower values of b one may expect a lower reliability

Another approach to the same type of statistic gives a formula for the standard error of measurement. If a large number of measures of a , say $a_1, a_2, a_3, \dots, a_n$, and corresponding measures $b_1, b_2, b_3, \dots, b_n$ are obtained for the same person by a series of n measurements, a set of values r_i will also be obtained. The standard deviation of this set is by definition the standard error of measurement of the ratio. From (2),

$$\sigma_{r_i}^2 = \frac{M_{a_i}^2}{M_{b_i}^2} (v_{a_i}^2 - 2r_{a_i b_i} v_{a_i} v_{b_i} + v_{b_i}^2) \quad (5)$$

If one assumes that errors in a are independent of errors in b , when the same person is tested repeatedly,

$$\sigma_{r_i}^2 = \sigma_{\text{meas.}}^2 = \frac{M_{a_i}^2}{M_{b_i}^2} \left(\frac{\sigma_{a_i}^2}{M_{a_i}^2} + \frac{\sigma_{b_i}^2}{M_{b_i}^2} \right) \quad (6)$$

or, if s is used as a symbol for the standard error of measurement,

$$s_i^2 = \left(\frac{M_a}{M_b} \right)^2 \left(\frac{s_a^2}{M_a^2} + \frac{s_b^2}{M_b^2} \right) \quad (7)$$

where s_a and s_b can be obtained by the usual methods. This reduces, using the identity $s = \sigma \sqrt{1 - r_{ii}}$, to (4). It must again be stressed that this formula is valid only where $\frac{\sigma_b}{b}$ is small enough that powers may be disregarded.

Since the formulas (4) and (7) are valid for some sets of scores and invalid for others, some procedure must be developed to determine where the formula is applicable. A useful test to determine whether the formula applies to any set of scores is to obtain values for M_i and σ_i empirically. If the values are close to those found by formulas (1) and (2), the assumption may be considered reasonable in this case; if discrepancy appears, the formula should not be used.

With such a score as Per Cent Inconsistent on 1.41, the assumption may hold for high values of b , but not for scores based on small values of b (Total Reasons). In this case it is possible to compute by formula the reliability of those scores where the assumption applies, but not for cases where the denominator is small. To determine the range where the formula can be used, the following procedure has been found efficient.

1. A scatter diagram of a against b is made for the sample.
2. An estimate is made that the assumption will hold for values of b greater than a certain value, say b' . For all cases where b is equal to or greater than b' , the standard deviation and mean of b , and r_{ab} , are computed from the appropriate rows in the scatter diagram.
3. In a separate scatter diagram, t is plotted against b , using the same class intervals for b as before. For all of the cases which fall in rows so that $b \geq b'$, σ_t and M_t are computed by the usual method.
4. Using formulas (1) and (2), σ_t and M_t are computed from the data obtained in step (2). If these values are equal, or virtually so, to those obtained empirically for the same population in step (3), the assumption that the variation of b is small compared to b itself is probably justified for $b \geq b'$.
5. If the values from steps (3) and (4) are equal, it is possible that the assumption holds for a value $b'' < b'$. If the values from steps (3) and (4) are not equal, it is necessary to test a value $b'' > b'$. In either case, a new hypothesis is made, that the assumption holds if $b \geq b''$. Using the same scatter diagrams, values are calculated for the means, standard deviations, and r_{ab} for cases at or above b'' . This is a comparatively simple step, as most of the previous computation can be used again. Again, the values of M_t and σ_t obtained by formula are checked against those obtained empirically. By a repetition of this process, it is possible to determine the smallest value $b^{(n)}$ of b for which the statistics derived empirically and by the formula are equal within prescribed limits of accuracy. It is probably unnecessary to compare the means, as a check between the estimated and empirical standard deviations should be an adequate test, since it requires little additional work to check means also, it is probably wise to do so.

RELIABILITY OF RATIO SCORES

Having identified the range of b for which the assumption holds, one may compute the reliability or standard error of measurement by formula (4) or (7). Except for r_{aa} and r_{bb} , the statistics which enter this equation have already been computed in the steps above. In many cases it is most simple to compute r_{aa} and r_{bb} by the split-half method. If the Kuder-Richardson method is used, it is necessary to make an item analysis of those papers whose b -values are sufficiently high, separate from such an item analysis of all papers as would ordinarily be made for other purposes. It is possible to plan the item analysis in advance, ranking papers in the order of their b -scores, so that the Kuder-Richardson method may be applied to a portion of the population economically.

Summary

Methods appropriate for computing the reliability of ratio scores have been discussed. They are

(1) The retest method, which requires construction of a parallel form for greatest meaning. The necessity of a second testing makes this inapplicable in many situations. A coefficient so obtained assumes that the reliability of all scores in a group is the same.

(2) The Spearman-Brown formula, applied to the correlation between scores based on a splitting of the test into two parts. This is generally invalid for ratio scores.

(3) The Kuder-Richardson formula, which may be used only where the denominator of the ratio is a constant.

(4) The Holzinger formula for the reliability of ratios, valid only if the variation of scores in the denominator is small compared to the mean of the denominator for the group. A related formula for the standard error of measurement developed in this paper is valid under the same conditions.

It was pointed out that ratio scores within the same

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

population, and even ratios which are equal in size, may not have the same reliability. The standard error of measurement increases as the denominator decreases. It follows that a single reliability coefficient for a ratio score is not meaningful, except for data where the variation in the denominator is small compared to the denominator.

GUIDING STUDENTS TO BECOME SELF-GUIDING

JOSEPH S KOPAS

Fenn College

TO adjust oneself to modern living requires a degree of personal development that informal, hit-and-miss efforts of the individual do not supply. Too many people are frustrated and unhappy because their preparation for adjusting themselves to our complex society has been only incidental. In this day and age a person requires training specifically directed at teaching him how to get the most out of life and the most out of himself. He must learn to appraise himself, to direct his personal development, to take advantage of opportunities for growth, and to evaluate his progress from time to time. The development of these skills should not be left to chance. They are far too important for that. An organized, formal program of training in self-guidance is needed.

In recognition of this need, a guidance program has been developed over a period of years at Fenn College. The use of evaluative procedures is an integral part of the program.

Fenn College utilizes the cooperative plan of education. The students are divided into two groups: one in class at the college, the other in full-time work off campus. At the end of each three-month period they alternate. The cooperative work experience received by the students is an important factor in the guidance program.

Objectives of the Guidance Program

At the time the guidance program was first considered in 1931 as an organized activity of the college, it seemed logical that the following objectives be kept in mind

- 1 *That the guidance program be an integral part of the college program.*

The guidance program should perform so essential a function that the contribution and progress of the guidance activities could best be judged by the progress made by the institution as a whole

- 2 *That the guidance program be centered on the normal student*

In too many cases, because of lack of time and personnel, only the maladjusted individuals in college get any real assistance from the guidance program and the normal student is, to a large extent, disregarded. By stressing the preventive as well as the adjustment phases of guidance work, and by developing techniques and methods which would be helpful to the normal students, it was hoped that the primary function of the guidance program would be to help the normal individual

3. *That all faculty members participate*

It was thought desirable to have every faculty member share in the program so that all students could be assisted properly. It was assumed that every faculty member could do some formal guidance work and that in doing his part he could, if given proper help, progressively qualify himself to do more and to do it better. Furthermore, it was assumed that the counseling experiences could help him to become a more effective teacher. Therefore, participation was expected to be of personal value to the instructor who shared in the program

GUIDING STUDENTS TO BECOME SELF-GUIDING

Organization of the Guidance Program

During the past ten years, a guidance program was evolved which is in line with the above objectives. A list of the features of the program, with a brief description, follows

1 The guidance program starts with the student

It helps him face as much of the responsibility for directing, motivating, and appraising the personal development as is educationally desirable. It expects and requires progressively greater assumptions of that responsibility as the student becomes more experienced and more capable.

2 Each instructor assumes a share of the responsibility in the guidance program as a general counselor

Each instructor acts as a general counselor for at least 10 students. The counselor is the institution's representative who assumes the responsibility of seeing that everything within the power of the college is done to help the student carry out his program of development to a successful conclusion. All formal guidance work, except in abnormal or unusual cases, is carried out through the counselor.

3 A guidance specialist is provided to serve as a supervisor of the counselors

His responsibility is to organize the program, to select and develop techniques, and to provide leadership and direction to the guidance program.

4 Each freshman student is enrolled in a group guidance class, called the Orientation Class

This feature will be described in detail later in this article.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

- 5 *Faculty case board conferences are held at the end of each quarter*

At these conferences the work of all students is reviewed and the progress and difficulties discussed. This activity is a very effective part of the guidance program and provides very good in-service training media in guidance techniques and methods for counselors.

- 6 *Various useful data are collected and made available to both the student and the counselor*
These data include test results, records, and other vital information.

- 7 *A clinic is provided to deal with problem students.*

Specialists participate in this clinic. This feature is still in its early stages of development.

The Record and Planning Folder

Space does not allow a detailed description of each feature of the program. However, the Record and Planning Folder and the Orientation Class, because of their uniqueness and importance in the guidance program, will be discussed here in greater detail.

The common practice is for colleges to keep records of the student's plans, achievements, and experiences. This practice takes care of the administrative needs, but does not give the student an opportunity to learn how to keep his own records. Planning requires that reliable information be gathered, organized, and used. For that reason it is important that the student learn how to keep records as a part of his training in self-guidance.

Three years ago a group of students in an orientation class decided to do some pioneering work in the area of personal record keeping. The form developed was called the Record and Planning Folder. The students found the record very helpful and were quite enthusiastic about its value. The folder provided a convenient method of

GUIDING STUDENTS TO BECOME SELF-GUIDING

gathering and organizing information necessary for effective self-guidance

Most of the information used in the folder was already available but seldom organized by the students. The following items were placed in the Record and Planning Folder on specially prepared forms during the freshman year.

1. *Personal history*

This section includes personal data, such as date of birth, father's and mother's names, occupations, nationality, names and ages of brothers and sisters in the family, and the student's employment experience prior to entrance in college.

2. *Autobiography*

The autobiography is a report of approximately 1,500 words containing the highlights of the student's history.

3. *Summary of high school record and experiences*

This summary includes all the subjects taken by the student and the grades, listed chronologically, as well as the student's rank in class, honors and scholarship, special courses, extracurricular activities, and his appraisal of his high school experiences.

4. *Entrance test results*

Each freshman undergoes a two-day testing program as part of the freshman week activities. The tests are of the type which the average instructor and student can understand and use. For that reason, the information is made available to both the student and the counselor. Areas of testing and the tests given are as follows:

General Ability Tests.	A C E Psychological Examination and Otis Mental Ability Test
General Background Tests	Cooperative General Achievement Tests in Natural Science, Social Science, Mathematics, and English

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Training in Special Subjects	Iowa Placement Examinations in Mathematics and Chemistry Training
General Aptitudes	Iowa Placement Examinations in Mathematics, English, and Chemistry Aptitudes
Vocational Interests.	Strong's Vocational Interest Test, using the modified scoring
Special Skills	Reading Test
Personal Characteristics	A battery of tests developed by the author

5 *Report on the tentative plans for the school year*

This report covers the general plans for personal development that have been worked out with the the student's counselor. In addition to the academic plans, they include plans for work experience, extracurricular activities, social development, and community and religious activities.

6 *Scholastic record for the freshman year*

Subjects taken and the grades received during each quarter, point averages and rank in class, and the student's appraisal of the work done each quarter are included

7 *Cooperative work experience*

Freshmen normally start on their cooperative work experience at the end of the third quarter. Each student is required to write a report about this work experience. The highlights of this report, experience received, earnings, the employer's evaluation of the student's work, and the student's appraisal of the work experience, are placed in this section

8 *Record of unusual experiences and opportunities utilized*

This record includes extracurricular and community activities in which the student engaged, worthwhile social and religious experiences, as well as any unusual activities

9 *Report on life philosophy*

As a part of the Orientation Class activities the student states his philosophy in terms of a

GUIDING STUDENTS TO BECOME SELF-GUIDING

pattern of beliefs. He inserts this in the Record and Planning Folder.

10 *Appraisal and evaluation of progress and experience during school year*

In this section the student records the highlights of the appraisal he and his counselor have made of his progress and difficulties, and modifications of plans made as the year progressed.

Provisions in the Record and Planning Folder are made for the following information for each succeeding year:

- 1 Addition to the autobiography
- 2 A report on tentative plans for each year (Made prior to registration)
- 3 Scholastic record
- 4 Cooperative work (work experience record)
- 5 Record of unusual experiences and opportunities utilized
- 6 Appraisal and evaluation of progress during the school year (Made at the end of each school year)

The Record and Planning Folder might very easily become one of the most significant features of the guidance program because it is helpful in so many different ways to the student himself and to the faculty members who deal with him. This coming school year every student will maintain the Record and Planning Folder as a part of his own guidance efforts.

The Orientation Class

The purpose of the Orientation Class is to help students intelligently plan and carry out a program of personal development that will lead to successful adjustment in major areas of adult responsibilities. It is an activity which provides opportunity for formal training in the process of self-guidance.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

The students and instructors jointly and mutually assume responsibility for planning, organizing, conducting, and evaluating the Orientation Class activities. Consideration of problems of personal adjustment and personal development makes up the program. Three weeks are spent in planning the program, seven weeks in carrying it out, and one week in evaluating it.

The general theme of the course is "Learning to Live in Our Modern Complex Society." The areas of adult activities which have been chosen to constitute the points of emphasis of the course are as follows: Learning to live with (1) one's self, (2) others, (3) one's job, (4) one's government, (5) one's estate, (6) one's culture, and (7) one's family.

Each student as a member of a committee helps plan a program consisting of three one-hour sessions in one of the above seven areas that he chooses, and then helps conduct the class activities. At the end of the week, each student turns in a report which includes his objectives, his problems, and his plans for personal development in the particular area under discussion. By the end of the quarter each student has written in detail a report containing seven sections stating how he intends to utilize the opportunities for personal growth to be found in college, in his cooperative work, and in community activities.

The advantages and importance of such a group guidance activity are readily seen. In the first place, the students are oriented into the major areas of adult life activities and responsibilities, in the second place, they are given a demonstration, through group thinking, of how a student, by means of choice and planning of activities, learns to assume more responsibility, exercises a greater use of his intelligence, attains greater control of his behavior, and is able to evaluate his experiences more meaningfully, than the student who merely drifts with the current of events in college. Finally, in a friendly, informal atmosphere, a stimulating environment is pro-

vided for the student so that he gets a good start on his self-guidance program

Difficulties Encountered

Difficulties that are encountered in the guidance program are common and familiar to all guidance workers

One of these difficulties is that of making the term "learning self-guidance" more concrete, both as to what is to be learned and how it is to be practiced. We have found that a practical approach to the difficulty is to limit the formal training the students are to receive in the area of self-guidance to the following three aspects:

- 1 The development of a dynamic outlook on life—as a source of direction and motivation.
- 2 The acquisition of a basic knowledge of the planning process—as a means of organizing and directing one's efforts
- 3 The maintenance of a record—as a means of evaluating one's progress and as a means of interpreting one's efforts to others

Each year that the student is in college he has an opportunity to formulate and discuss with his counselor the objectives of personal development he wishes to achieve during the year and plans for achieving those objectives, as well as any modifications of his plans or objectives made during the year. At the end of the year, he and his counselor evaluate the progress made. If the student follows this procedure each year he is in college, he will have practiced self-guidance in a very effective and worthwhile way.

Another difficulty faced is that of getting the instructors, who are busy and often not too interested or qualified in guidance work, to put the necessary effort into the job. We have tried to minimize the "too busy" problem by (1) making the student more active in the program, (2) making the information about students quickly and easily available and usable, (3) distributing the task equally among all the instructors.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

The "not interested" problem was tackled by (1) expecting and giving the instructors an opportunity to participate on the theory that participation stimulates interest; and, (2) getting them to see that the guidance program contributes to the improvement of their main function, which is teaching.

Finally, the "not qualified" problem was handled by (1) providing in-service training for the instructors, and, (2) simplifying the techniques and devices to a point where the average teacher can use them.

Another difficulty is that of overcoming student indifference and the tendency to drift. Guidance implies motion. Self-guidance, therefore, implies self-propelled motion. It is absolutely essential in the guidance program that the student take the initiative in developing himself. The Orientation Class, the counseling system, and the Recording and Planning Folder help motivate the student to take the initiative, rather than to sit back and drift with the current of events.

It is not possible to make an evaluation at this time of the complete effectiveness of the guidance program. A survey of the results up to this date would show that about 20 per cent of the students and 40 per cent of the faculty members are doing a good job of their respective parts of the program. Almost one-third of the faculty and one-fourth of the student body are not functioning very effectively. The remainder are doing just a fair job. Admittedly, progress has been slow. But the participants are becoming more and more interested as time goes on, and the program appears to be growing in effectiveness. Progress should be more rapid within the next five years now that all the features described are included in the program.

AN ATTEMPT TO MEASURE SCIENTIFIC THINKING

MAX D. ENGELHART AND HUGH B. LEWIS

Chicago City Junior Colleges

POSSIBLY the most challenging problem facing those engaged in the construction and use of objective tests is the creation of exercises which will require the functioning of abilities transcending memory. The series of exercises presented in this paper may not deserve the label of a test of the ability to think scientifically. It seems justified, however, to present the series in the hope that the form of the exercises may suggest to other and more ingenious test makers improved means of measuring abilities which are among the universally recognized goals of science instruction.

The series of exercises given here follows in its organization the steps often regarded as the essence of the scientific method. One would be naive to believe that scientific problems are always solved in just these steps, or that the processes involved in their solution may not be more complex. On the other hand, the use of the stages represented in these exercises may be appropriate in testing students. Although the proper function of a test is measurement, it may still be legitimate to recognize the function of motivation, and exercises of the type classified may also accomplish the purpose of engendering in the minds of students a belief that knowledge of the scientific method is important.

When the exercises were constructed, it was felt essential to present certain introductory statements descriptive of the scientific method and of the phenomenon with which the problem to be solved is concerned. The distinc-

tion between the directness and indirectness of the contribution of a datum in determining the truth or falsity of an hypothesis may be somewhat artificial. It is possible that the use of three categories rather than five would be justifiable. One might argue, however, that scientific thinking does involve this evaluation of relevancy of data, and that the more direct the contribution, the greater is the dependence which may be placed upon the data.

When the content of the exercises was selected, an effort was made to present a phenomenon which in most respects would be novel to the students—that is, the phenomenon would be novel, but the concepts basically involved would relate to subject matter or principles with which the students had had some experience. The exercises on the operation of the radiometer were developed from a series of exercises of a somewhat different type which were written by Dr. C. E. Ronneberg of Herzl Junior College for the January, 1939, physical science comprehensive examination. It is possible to select other phenomena for which problems can be stated, and to construct similar exercises. There is, of course, no necessity to restrict such exercises to the field of physics. The particular phenomenon and exercises presented here were not an altogether appropriate selection so far as the group tested was concerned, since the level of difficulty was too great.

The series of exercises was included in a test administered to students entering the Chicago City Junior Colleges who wished to enroll in the second, rather than in the first, semester of the physical science survey. The exercises and their introductory materials are reproduced below.

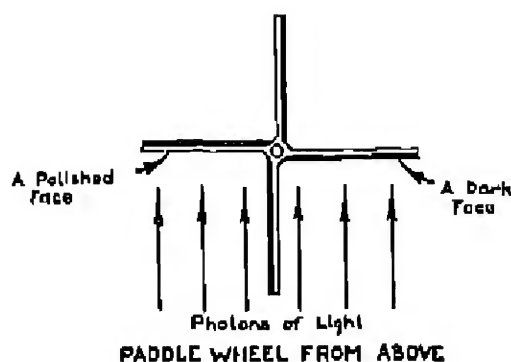
A scientist, when confronted with a problem, formulates hypotheses which represent tentative solutions to the problem. He then collects data which may support or disprove his hypotheses. Finally, on the basis of the data and the hypotheses thus tested, he derives a conclusion which constitutes his answer to the problem.

AN ATTEMPT TO MEASURE SCIENTIFIC THINKING

The following exercises represent an effort to test your ability to do scientific thinking. You are to test certain true or false hypotheses, and to evaluate certain general conclusions. Assume that each item of data below each hypothesis is a true statement and *may* directly or indirectly help to prove an hypothesis true or false.

If the application of the item of data requires only one step to prove the truth or falsity of an hypothesis, then the item is a *direct* help. For example, the temperatures of water boiling on a given mountain and at sea level would represent *direct* evidence of the falsity of the hypothesis "water boils at a higher temperature on a mountain than at sea level."

If the application of the item of data requires more than one step to prove the truth or falsity of an hypothesis, then the item is an *indirect* help. For example, the item "water in a container that can be evacuated will boil at 100m temperature" *indirectly* helps to prove the falsity of the hypothesis "water boils at a higher temperature on a mountain than at sea level."



A number of years ago Sir William Crookes perfected an instrument which always intrigues people, whether laymen or scientists. This is the radiometer, a device consisting essentially of a paddle wheel which is free to rotate in a horizontal plane within a partially evacuated glass bulb. One side of each paddle is brightly polished,

while the other side is coated with lampblack. As soon as the device is placed in the sunlight, the little paddle wheel starts to spin rapidly. It continues to spin until the device is again placed in the dark.

PROBLEM How does sunlight cause the paddle wheel to rotate?

Below are given a series of hypotheses, each of which is followed by numbered items which represent data. After each item number on the answer sheet blacken space.

- A if the item directly helps to prove the hypothesis true,
- B if the item indirectly helps to prove the hypothesis true
- C if the item directly helps to prove the hypothesis false
- D if the item indirectly helps to prove the hypothesis false.
- E if the item neither directly nor indirectly helps to prove the hypothesis true or false

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

HYPOTHESIS I In a partial vacuum the paddle wheel rotates because of the impact of photons of light

- 128 Scientists now believe that light has both corpuscular and wave characteristics
- 129 In a very high vacuum the bright faces of the paddle wheel turn slowly away from the light, while the black faces turn toward the light
- 130 Light travels at the rate of 186,000 miles per second
- 131 In a partial vacuum the black faces of the paddle wheel turn away from the light, while the bright faces turn toward the light
- 132 Light travels at a slower speed in glass than in air or in a vacuum
- 133 After *this* item number on the answer sheet blacken space *A* if Hypothesis I is true, or space *B* if it is false

HYPOTHESIS II A paddle wheel on which all of the faces are bright or all are black will not rotate

- 134 The black faces of paddles absorb energy from light to a greater extent than the bright faces of paddles.
- 135. Rotation is due to force of impact If all paddles are the same on both sides, either all bright or all black, the turning forces would cancel
- 136. More photons rebound from bright faces than from dark faces
- 137 In a partial vacuum, air molecules are constantly hitting the paddles
- 138 Photons are hitting the sides of the paddles which face the light
- 139 After *this* item number on the answer sheet blacken space *A* if Hypothesis II is true, or space *B* if it is false

HYPOTHESIS III Rotation in a partial vacuum of the paddle wheel is due to the greater force of rebound of air molecules from the black faces than from the bright ones

- 140 The bright faces remain cooler than the dark faces, since they reflect more light
- 141 In a partial vacuum and in the dark the paddle wheel will rotate when exposed to invisible infrared rays from a warm flatiron

AN ATTEMPT TO MEASURE SCIENTIFIC THINKING

- 142 The black faces of the paddles become warmer than the bright faces, since they absorb more light
- 143 Air molecules adjacent to the warmer black faces rebound from these faces with greater energy than from the cooler bright faces
- 144 In a very high vacuum and in the dark the paddle wheel will rotate slowly if invisible rays from a cathode tube are directed toward it
- 145 After *this* item number on the answer sheet blacken space *A* if Hypothesis III is true, or space *B* if it is false

Below are five conclusions After each corresponding number on the answer sheet blacken space

A if in your judgment the conclusion is the best answer to the problem

B if in your judgment the conclusion is neither the best answer nor the least satisfactory answer to the problem (Three conclusions should receive this mark.)

C if in your judgment the conclusion is the least satisfactory answer to the problem

- 146 The paddle wheel of the radiometer rotates, because air molecules move with greater energy when heated by energy from sunlight or from infrared rays from a flatiron
- 147 Air molecules rebound with greater force from the bright faces, which reflect more light energy Photons rebound from dark faces to a greater extent than from bright faces The turning forces thus created cause black faces to rotate toward the light in a partial vacuum and away from the light in a very high vacuum
- 148 The paddle wheel of the radiometer rotates, because photons of light strike air molecules with greater energy when adjacent to the dark faces than when adjacent to the bright faces
- 149 The fact that a radiometer will operate in either a partial or a very high vacuum demonstrates that it is not essential that air molecules be present in order to cause rotation
- 150 Air molecules rebound with greater force from the black faces, which absorb more light energy than the bright faces Photons rebound from bright faces to a greater extent than from dark faces. The turning forces thus created cause black faces to rotate away from the light in a partial vacuum and toward the light in a very high vacuum

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

In the table below are presented the proportions of correct response and correlations of each of the items with the total score on the test and with the total score on the part, i e , the total score on the 23 scientific thinking exercises. These data are based on an analysis of the answer sheets of a random sample of 200 cases. The selection

TABLE 1
ANALYSIS OF ITEMS

Item No	Key	Item Difficulty	Item-Test Correlation	
			Correlation with Total Score on Entrance Test	Correlation with Total Score on Part
128	E	17	25	35
129	D	09	15	20
130	E	55	58	61
131	C	18	15	27
132	E	42	37	55
133	B	21	18	30
134	E	12	19	32
135	A	60	41	41
136	E	18	15	33
137	E	28	40	55
138	B	23	28	40
139	A	65	40	38
140	B	22	00	27
141	E	18	27	27
142	B	20	08	36
143	A	41	28	33
144	E	24	41	48
145	A	33	09	29
146	B	41	38	38
147	C	15	20	20
148	B	50	35	45
149	B	39	09	19
150	A	39	19	17

was made from the answer sheets of all of the students taking the test on entrance into the junior colleges. The reliability of the series was found to be .72 by means of a Kuder-Richardson formula. The series of exercises correlated .64 with the total score on the entrance test. Seventy-five of the other exercises were factual, multiple-answer items dealing with high-school physics and chemistry, and 57 were true-false items pertaining to several passages selected from advanced texts in the physical science field. These latter exercises emphasized aptitude more than training in that they were essentially a reading test in the field of physical science.

AN EVALUATION OF TECHNIQUES OF MEASURING VISUAL ACUITY AT THE COLLEGE LEVEL

FRANCES ORALIND TRIGGS

University of Minnesota

KARL E. SANDT, M.D.

University of Minnesota Health Service

THE UNIVERSITY of Minnesota Health Service and the University Testing Bureau have been cooperating on an evaluation of the Betts Ophthalmic Telebinocular Test to determine whether it is a valid screening test of visual acuity for use with college students.

The problem of determining what students should and what students should not be referred to an eye specialist is a real one because many health services do not have such doctors on their staffs and students, many of whom cannot afford it, must pay for such service individually. If there is an eye specialist on the staff of the health service, it is difficult for him to give each student individual attention. The students who come voluntarily to him to be examined may be the very ones who do not need an examination, and those who do need it may never get it, for often a student himself does not know when his eyes need attention.

The plan of the research was this. The Betts Telebinocular Examination was included as a part of the diagnostic reading test battery which is given by the University Testing Bureau in cases of suspected reading difficulties. At the time the student took the Telebinocular Test, he was given a note addressed to Dr. K. E. Sandt at the Health Service asking for a complete eye examination. This note was an indication to Dr. Sandt that the student was to be included in the research. It

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

was intended to have complete data from these tests on 100 students, but when the data from the records on the Telebinocular and from the Health Service records were tabulated, it was found that data were complete for only 87 students. The measures on which data were available from both sources were visual acuity in the right and left eye separately, exophoria and esophoria near and far, and hyperphoria near and far. If a student wore glasses he was supposed to be tested with and without glasses. If the student wore glasses to the examination at the Health Service and the examination at the University Testing Bureau, the data with glasses were always used. If data were available only from one examination without glasses, the data from both services without glasses were used.¹

In evaluating these data, certain facts should be kept in mind. The Betts Telebinocular² purports to screen out students having measurable visual defects serious enough to be considered for correction by an ophthalmologist. In a letter to the University Testing Bureau from the Bureau of Research of the Keystone View Company dated April 22, 1940, the following bases for referral were given: "(a) You need have no hesitation about referring a patient who fails on any part of Test 3, provided that if there has been a failure on B or C with both eyes open you make the test by occluding the eye not being tested. If there is still a failure the patient should be referred. (b) Test 4 is seldom failed but if it is failed, without question the patient needs attention. (c) The failure of Test 5 alone is not a warrant for referring the patient, but if there have been other failures, particularly in Test 2 and 6A, there is no question but what there is poor eye co-

¹It should be remembered when interpreting these data that there is no indication as to whether or not this group of students is a selected sample of the whole student body as far as visual acuity is concerned. There is no reason to believe that they would be a selected sample on the criteria of visual acuity just because they are on reading skills, for it has been shown that there is no consistent relationship between these two factors.

²For complete description of the instrument, see Emmett Albert Betts, *The Prevention and Correction of Reading Difficulties* (Evanston, Illinois: Row, Peterson and Company, 1936), pp. 327-50.

TECHNIQUES OF MEASURING VISUAL ACUITY

ordination (d) With high school and university students who complain of discomfort in reading, failures of Test 6B and 7 taken together are indicative of near point trouble, and they should have attention " It is upon these bases that referral was determined by the University Testing Bureau for this study The ophthalmologist is trained to determine visual defects and decide whether they are serious enough for correction, thus it may be seen that there is some overlap of service but no overlap of responsibility, the final decision always lying with the ophthalmologist as to whether correction shall be given In the light of these stated purposes of the Betts test, it would seem that the following questions might be helpfully answered

1 Will the Betts test screen out for referral to the oculist a large number of students in whom the oculist will find deficiencies serious enough for correction?

2 Will the Betts Telebinocular refer a large proportion of students whom the oculist finds to have no measurable eye difficulty?

3 By comparing the oculist's and Betts' records, on individual tests, are all of the Betts measures equally satisfactory?

4 On what tests of the Betts Telebinocular are referrals made most often? Can a better basis of referral on the Betts Telebinocular be found than the one furnished us by the Keystone View Company?

The question which always arises in research of this kind is whether the tests measure what they purport to measure and whether, if they were administered a second time, they would give the same results as they did the first time These questions have never been finally answered for either of the tests used in this study It may be that they never can be answered, for in both cases the results are dependent upon the physiological status of the individual being tested The relationship of the effects of fatigue, light, and other factors upon different individuals may vary so greatly that a constant score may not

be possible. Or it may even be that there are still to be discovered better ways of diagnosing visual anomalies.

In this study, the extent to which the two tests agree on diagnosis is indicated, but neither test is assumed to give a perfect diagnosis. However, because prescriptions are finally made by the oculist, the extent to which the Betts would refer to the oculist those people found by him to need correction is pointed out.

The following evaluation of data is presented in answer to these four questions:

1. Of the 87 students included in this study, 13 were given glasses by the oculist, 11 were given prescriptions to correct measurable physical eye defects, and two students were given glasses merely to improve comfort while reading rather than to correct measurable eye defects.

Of the 11 students given prescriptions to correct measurable physical eye defects, all would have been referred on the criteria of referral sent us by the Keystone View Company. The remaining two students would not have been referred on the basis of these criteria. Thus it will be seen that all students found by the oculist to have measurable physical eye defects would have been referred for complete examinations as a result of the Telebinocular Test.

2. Of the 87 cases which had both the Betts test and an ophthalmic examination, 46 would have been referred by the Betts test to the oculist on the basis of the criteria furnished us by the Keystone View Company. Of the 46 students who would have been referred by the Telebinocular Test, only 11, or 24 per cent, had defects serious enough to be corrected by glasses. However, it should be remembered that while only about 53 per cent of the group would have been referred to the oculist for complete testing, 100 per cent would have had to be tested, had no pre-test been given. While it might be desirable to have a more rigorous screening test, it is certainly worth-while to save the oculist from having to examine almost half of the students.

TECHNIQUES OF MEASURING VISUAL ACUITY

3 The data which bear on this question are presented in Table 1. Data were complete from both sources for 87 students on visual acuity for the right and left eyes. For the right eye, we find that 20 students, or 23 per cent, failed the Betts test and the oculist's test, 48, or 55 per cent, passed both tests. In other words, it was found that the oculist's diagnosis and the Betts' diagnosis agreed on 78 per cent of the cases. Six students, or 7 per cent, failed the Betts test but were found satisfactory by the oculist, and 13, or 15 per cent, passed the Betts test but failed the oculist's test. It should be remembered that of these 13 who passed the Betts test but failed the oculist's test, the defect found by the oculist was in no case considered serious enough for correction.

TABLE 1
COMPARISON OF RESULTS BETTS TESTS AND OCULIST'S TESTS FOR EIGHTY-SEVEN SUBJECTS

Failed Betts and Oculist's				Passed Betts and Passed Oculist's			
Right Eye		Left Eye		Right Eye		Left Eye	
No	%	No	%	No	%	No	%
20	23	13	15	48	55	54	62
Failed Betts and Passed Oculist's				Passed Betts and Failed Oculist's			
Right Eye		Left Eye		Right Eye		Left Eye	
No	%	No	%	No	%	No	%
6	7	10	11.5	13	15	10	11.5

For the left eye, data were again complete for 87 cases. We find that 13 students, or 15 per cent, failed the Betts test and the oculist's test, 54, or 62 per cent, passed both tests. Thus the oculist's and the Betts' diagnosis agreed on 77 per cent of the cases. Ten students, or 12 per cent, failed the Betts test but were found satisfactory by the oculist, and 10, or 11 per cent, passed the Betts test but failed the oculist's test. It should be remembered that of those 10 who passed the Betts but failed the oculist's tests, the defect found by the oculist was in no case considered serious enough for correction.

On the measure of vertical imbalance (hyperphoria) far point on the Betts test, none of the 72 cases on which data are complete for both measures failed the test. (Two stu-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

dents who were included in the study, but for whom no oculist measure of vertical imbalance far point was available, failed this test) The oculist found that 10 students, or 14 per cent, of the 72 students for whom data were complete on both measures had vertical imbalance and 62, or 86 per cent, did not

The Betts test for vertical imbalance near point did not identify any member of the group with this difficulty, but the oculist found that seven, or 10 per cent, of the 72 did have vertical imbalance and 65, or 90 per cent, did not have vertical imbalance near point These data would seem to indicate that this test is of questionable value for use with college students

On the lateral imbalance near point, data are complete for 72 cases Of the 72 cases, four students, or six per cent, failed the Betts and the oculist's tests, two, or three per cent, failed the Betts and passed the oculist's test, 18, or 25 per cent, passed the Betts and failed the oculist's test, and 48, or 66 per cent, passed the oculist's test and the Betts test Thus it will be seen that the two measures agreed in 72 per cent of the cases

For lateral imbalance far point, no students of the 72 who were found to be unsatisfactory by the oculist failed the Betts, but two students, or three per cent, failed the Betts who were found to be satisfactory by the oculist; four, or five per cent, passed the Betts test who were found to be unsatisfactory by the oculist, and 66, or 92 per cent, were found to be satisfactory on both measures For lateral imbalance far point, the two measures agreed in 92 per cent of the cases

This evaluation of tests would lead us to say that, of the parts of the Betts tests studied, the one found to be most valuable for referral of college students for a complete eye examination is the one of visual acuity for the right and left eye There is not complete agreement between this test and the oculist's measurements, but it does agree 78 out of 87 times for the right eye and 77 out of 87 times for the left eye, and where it does not agree the oculist found no situation to exist which was serious enough for correction This finding raises the question as to whether another measure of visual

TECHNIQUES OF MEASURING VISUAL ACUITY

acuity not requiring expensive apparatus would serve as satisfactorily. Only one other measure of visual acuity was given these students. When students enter the University of Minnesota they are required to have a physical examination at the Health Service. As a part of that examination the eyes are checked by use of the Snellen Chart³. The record of these examinations was on file at the Health Service and has been tabulated for consideration here.

There was a record of a Snellen examination for all the 87 students included in this study. On the basis of that examination eight students would have been referred to the oculist for complete examination. For this study it is important to determine whether these are the same students referred by the Betts test and also to ascertain in how many cases the students referred were found by the oculist to have a defect serious enough to warrant a prescription.

Of the eight students referred by the Snellen test, six would also have been referred by the Betts. As has been stated, 13 of the 87 students were found by the oculist to have defects serious enough to be corrected by glasses. Of the 13 given glasses, the Snellen Chart examination would have referred only two to the oculist. These data would seem to indicate that on measures of visual acuity the Betts test is superior to the Snellen Chart in referring students with actual difficulty to the oculist for complete examination.

4. It will be remembered that the Keystone View Company gave four standards for referring a student to the oculist on the basis of the Betts test. These were

- (a) Failure on *test 3 or any part of test 3* (visual acuity)
- (b) Failure on *test 4* (vertical imbalance)
- (c) Failure on *test 5* (coordination) with failure on *test 2* (distance fusion) and failure on *test 6A* (lateral imbalance far point)
- (d) A *complaint of discomfort in reading* with *test 6B* (lateral imbalance near point) and *test 7* (fusion at reading distance)

³*Ibid.*, pp. 149-51

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Of the 46 students who would have been referred to the oculist on these standards of referral for the Betts test, 43 are identifiable by their failure on the first criterion, i.e., failure on one or more of the visual acuity tests. Two of these students also failed the vertical imbalance far point test. Two of the students who would have been referred failed on both criteria one and four. Two failed criterion four only and would have been referred on it alone. One student failed and would have been referred on criterion three alone.

Thus, examination of our data would indicate that referral on the basis of criteria one, three, and four and the record of a complaint of discomfort would have referred all students given prescriptions by the oculist to correct measurable physical eye defects. The number referred would have been 46 and it would have included the same students referred by the four criteria given by the Keystone View Company. It is questionable whether the test of vertical imbalance adds anything essential to this series when used as a screening device at the college level.

On the basis of the data just presented it does seem that the Betts Ophthalmic Telebinocular Test can be used satisfactorily by colleges as a screening device for referral of cases to the oculist. The Betts test also stands up more satisfactorily than does the Snellen Chart examination (when given under the conditions described) as a measure of visual acuity.

These conclusions should be checked by repeating this study on another group of students, and they should be accepted only tentatively for situations other than those described in this study.

THE CONCEPT OF SCATTER IN THE LIGHT OF MENTAL TEST THEORY¹

MAURICE LORR

U S Civil Service Commission

RALPH K MEISTER

Mooseheart Laboratory for Child Research

THE CONFUSION and loose thinking among clinical psychologists concerning the basis and significance of scatter on scales of the Binet type suggest a re-examination of the concept of scatter in the light of the theory of psychological measurement

Theoretically, on mental age scales of the Binet type, items are arranged in the order of their difficulty, the easiest item first. In clinical practice these items are administered to a child in the same sequence. Groups of items, supposedly equal in difficulty, are allocated to each year level as representing the typical performance of individuals of the corresponding chronological age. The child is given increasingly difficult items until he reaches a point in the scale above which he fails. Actually, no such point exists. Instead, the child passes all tests at a certain level and continues with mixed successes and failures on to the next higher level until he fails all items presented to him in a given level. Such a spread of successes and failures over a number of mental year levels is called scatter.

Test theory indicates five possible bases for such irregularity of performance. First, scatter is a consequence of the lack of perfect correlation between test items resulting from the presence of error and from the low communality and

¹The authors wish to thank Dr. M. W. Richardson for his review of this article and Dr. M. L. Reymert for his interest and encouragement throughout.

high specificity of the items. The error, as Mosier (10) has shown, increases in the individual case with greater heterogeneity of items. Thus, an individual who passes one item at a given level may not necessarily pass a second item, either because the two items do not measure the same function or because of error involved in testing. Illustrative of this lack of perfect correlation between test items is the Cattell and Bristol study (4) in which a mean intercorrelation of $+ .32$ was found for seven Binet test items, Wright (15) found the mean intercorrelation on 31 items to be $+ .61$.

Secondly, there is the fact that the items are incorrectly allocated in the order of difficulty. This might be expected in view of the fact that Terman and Merrill (12), for example, although using curves of proportions-passing for each item in the preliminary grouping of items, had as their goal in the final grouping an I Q distribution with a mean of approximately 100. It is likely that such a procedure resulted in a grouping of items only roughly ordered as to difficulty. Therefore, although the grouping of test items is approximately in the order of their difficulty in the sense that an item at age four is definitely less difficult than one at age 10, nevertheless, in adjacent groups there are probably a great many inversions in difficulty. Thurstone's study (14) on the absolute scaling of Binet items shows that items at any particular age level vary considerably in difficulty, a finding contrary to the assumption of relatively equal difficulty among items at any one age level. He found, too, inversions in placement according to difficulty. Thus an item which is easier than another may be placed at a higher age level. In fact, allocation at the different age levels on the basis of difficulty is improbable since the items distribute very unevenly in absolute difficulty over the age levels. On the basis of Burt's data, Thurstone (14) says, "The test questions are more numerous at certain ages than at others. For example, there are 12 questions that scale at par between the ages five and six, whereas there are only four questions that scale at par between six and seven." Any kind of arrangement that requires

SCATTER IN MENTAL TEST THEORY

the same number of tests at each age level is unlikely to result in equal gradations of difficulty since the test items used do not scale into any such grouping

This fact of the incorrect allocation of test items at the various age levels is brought out by the findings of many other investigators. Cyril Burt (2) admits that no two editors agree about the correct order of mental age items and cites instances. Barber (1), on data for the revised Form L, found that five items were significantly easier and six items were significantly more difficult than their respective age placements would indicate. Likewise, Harriman (5) found (for the Revised Scale) that test items at year level XII seem to be more difficult than those at year level XIII, a fact which is confirmed by Carlton (3). Krugman (8) found that for New York school children, 25 of the items were incorrectly allocated.

Thirdly, scatter may be due to the lack of discriminatory power of certain items. A highly diagnostic item will discriminate sharply between individuals with ability above that required to respond correctly to the item and those individuals who lack such ability. For example, two items may be equal in difficulty (50 per cent pass at, say, age 11), but differ widely in diagnostic value. The psychometric curve for one item may extend over, say, years five to 14. The curve for another item that is more discriminatory will extend over a much smaller age range, such as eight to 12. This spread or scatter is manifestly a result of low diagnostic value. Thurstone (14) plotted curves of proportions-passing for a random selection of items from the Burt-Binet and found a "noticeable variation in the slopes of curves." It is probable, therefore, that some of the scatter found is due to these differences in the diagnostic value of the items which these examples illustrate.

A fourth source of scatter may be found in the fact that there is an increase in variability with an increase in absolute mean test performance (13). In other words, individuals apparently become more variable as they grow older. Since

individuals vary more among themselves, they must vary as to the number and types of items failed or passed. This can be easily seen, since the extent to which Thurstone's "primary mental abilities," for example, are present varies between individuals and within individuals, and the factorial composition of test items differs from item to item within the same age level. Thus an individual who passes one item at a certain age level may not pass another because the latter requires an ability which he does not have to the required degree. Again, an individual's failure on five items at a certain level is no sure indication that he will fail the sixth, since the sixth item may require an ability which he has to a marked degree. This tendency of scatter to increase with mean test performance or chronological age is checked in actual practice between the ages 10 and 12 (Reymert and Meister [11]) for within that age range the ceiling of the test begins to limit the amount of scatter possible.

A fifth possible cause of scatter is the presence of systematic errors in testing due to language handicaps, sensory defects, special training, lack of cooperation, and ambiguous scoring or instructions. Unlike chance errors which influence the results as often in one direction as in another and therefore can be assumed to cancel out one another, systematic errors have a consistent and cumulative effect that gives the results a constant bias. Obviously if a test shows constant bias for a given individual, it is unsuited for that individual, i.e., the individual differs sufficiently from the norm population to render the test inapplicable. If such a test is given, the person with language difficulty will tend to fail highly verbal items, the uncooperative individual will answer only those questions which he can be motivated to try, the individual with a slight hearing loss may miss the critical part of the question, etc. All of these factors tend to lower the basal level which *per se* gives a larger amount of scatter.

The uses to which measures of scatter have been put can now be critically examined in the light of the sources of scatter given above. Perhaps the first point that should receive

critical attention is that of the methods of measuring scatter, for there are a great many such methods and the amount of agreement among them seems to be inversely related to their number. Theoretically, if scatter represents the range of uncertainty of an individual's ability, the range within which lies the limit of his ability or the point beyond which he will fail all the items—a point which in actual testing practice does not exist—then scatter should be measured on a scale of absolute difficulty, as the distance between the easiest item failed and the most difficult one passed. Actually, as no such measure has ever been used, it is not surprising that, to quote Harris and Shakow (6), "research up to now has failed to demonstrate clearly any valid clinical use for such measures". These authors mention nine methods of measuring scatter and conclude that "at the present time it is impossible to state which is the best method of measuring scatter."

Now, in view of the uncertainty about measures of scatter, the uses to which they have been put become all the more questionable. It is common practice to use measures of scatter as indicative of epilepsy, psychosis, feeble-mindedness, emotional maladjustment, hypopituitarism, etc. Harris and Shakow's paper (6), in fact, deals with "the possibility of obtaining clinically significant information from numerical measures of scatter."

Studies which have indicated significant differences in the mean scatter for certain groups do not justify diagnosis of a particular condition in the individual case. And when one considers that for five papers that report such differences, there are four that do not (6), even differentiation for groups appears questionable. For indicating the type of condition in question, there certainly should be a more refined instrument than scatter on a test designed to measure intelligence.

From a consideration of the various bases for scatter, it is obvious that their influence is certainly greater than that of any chance errors. Therefore, in the light of these considerations, the use of scatter as even a crude estimate of the

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

measurement error of the test for an individual is not an acceptable procedure

It is also common practice among clinical psychologists to analyze the particular successes and failures of an individual to give a crude appraisal of his primary abilities as well as estimates of mental deterioration in these abilities. Such inspectional analysis is a questionable practice for the following reasons. First, the factorial composition of an item cannot be prejudged accurately. Such judgments are frequently in complete disagreement with factor analysis results. For instance, Wright (15) found that items involving repeating digits backwards did not necessarily involve memory ability but rather other factors. And yet how many failures on such items have been analyzed in reports as poor memory ability? Secondly, an item may be solved through the use of different abilities by different individuals and at different age levels. Thirdly, items may show fairly high loadings on more than one factor so that failure cannot be attributed to the lack of any one ability. Fourthly, such clusters of items have too low a reliability to have any real diagnostic value.

In summary, it is concluded that scatter is for the most part due to factors inherent in test construction plus certain systematic errors. In view of these facts, it appears that the possibility of ever securing clinically significant information from measures of scatter based on age scales in current use is slight indeed.

SCATTER IN MENTAL TEST THEORY

REFERENCES

- 1 Barber, E R "A Study of Scatter and the Relative Difficulty of Sub-Tests in the Revised Stanford-Binet," Master's thesis, University of Illinois, 1938
- 2 Burt, C "The Latest Revision of the Binet Intelligence Test," *Eugenics Review*, XXX 4 (1934), 255-60
- 3 Carlton, Theodore "Performances of Mental Defectives on the Revised Stanford-Binet, Form L," *Journal of Consulting Psychology*, IV (1940), 61-5
- 4 Cattell, R B and Bristol, H "Intelligence Tests for Mental Ages Four to Eight Years," *British Journal of Educational Psychology*, III 2 (1933), 142-69
- 5 Harriman, P L "Irregularity of Successes on the 1937 Stanford Revision," *Journal of Consulting Psychology*, III (1939), 83-6
- 6 Harris, A J and Shakow, D "The Clinical Significance of Numerical Measures of Scatter on the Stanford-Binet," *Psychological Bulletin*, XXXIV (1937), 134-50
- 7 Harris, A J and Shakow, D "Scatter on Schizophrenic, Normal and Delinquent Adults," *Journal of Abnormal and Social Psychology*, XXXIII (1938), 100-11
- 8 Krugman, Morris "Some Impressions of the Revised Stanford-Binet Scale," *Journal of Educational Psychology*, XXX (1939), 594-603
- 9 Mateer, Florence "Differential Syndromes in Stanford-Binet Failures" (Abstract), *Psychological Bulletin*, XXXVI (1937), 508
- 10 Mosier, Charles I "Psychophysics and Mental Test Theory Fundamental Postulates and Elementary Theorems," *Psychological Review*, XLVII (1940), 355-66
- 11 Reymert, Martin L and Meister, Ralph K "A Comparison of the Original and the Revised Stanford-Binet Intelligence Scales," *Educational and Psychological Measurement*, I (1941), 67-76
- 12 Terman, Lewis M and Merrill, Maud A. *Measuring Intelligence* Boston Houghton-Mifflin, 1937 Pp 461

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

- 13 Thurstone, L. L. "The Absolute Zero in Intelligence Measurement," *Psychological Review*, XXXV (1928), 175-97
- 14 Thurstone, L. L. "A Method of Scaling Psychological and Educational Tests," *Journal of Educational Psychology*, XVI (1925), 433-51
- 15 Wright, R. E. "A Factor Analysis of the Original Stanford-Binet," *Psychometrika*, IV (1939), 209-20

MEASUREMENT ABSTRACTS*

Adkins, Dorothy C "The Relation of Primary Mental Abilities to Preference Scales and to Vocational Choice" *Psychometrika*, V (1940), 316 (Abstract of a paper read at the September, 1940, meeting of the American Psychological Association)

Benge, Eugene J "Wanted More Logic and Less Guesswork in Hiring Salesmen" *Sales Management*, XLVIII, No 3 (1941), 18-20

Companies who have records for many employees, past and present, could profitably make an analysis of job requirements. It is suggested that a criterion of job efficiency be set up and employees classified accordingly. An outline, based on these data, for constructing a rating scale in terms of factors which can be elicited at time of application is presented. Minimum scores on the rating scale are to be established according to scores made by employees. It is emphasized that scores on such a rating scale should not be considered alone but only in connection with other sources of information in hiring employees. *D A Peterson*

Blackwell, A M "A Comparative Investigation Into the Factors Involved in Mathematical Ability of Boys and Girls" *British Journal of Educational Psychology*, X (1940), Pt I, 143-53, and Pt II, 212-22

A group of 100 boys and a group of 100 girls, ages ranging from 13½ to 15 years, were given 10 tests of "mathematical ability" including arithmetical reasoning, analogies, three "spatial tests," three "geometric tests," and a test of algebraic computation and reasoning. The intercorrelations for each sex were factored separately. Interpretations were attempted on the basis of the centroid matrices and on the

*Edited by Professor Forrest A Kingsbury

basis of an orthogonal rotation as designed to insure a general factor. Sex differences were found. "The results of the study seem to confirm the complex nature of mathematical ability." *Harold Bechtoldt*

Coombs, Clyde H. "A Factorial Study of Number Ability" *Psychometrika*, VI (1941), 161-89

In order to investigate certain hypotheses concerning the nature of number ability, and, secondarily, the nature of perceptual speed, a battery of 34 tests was given to 223 Chicago high school seniors and the data were factored by the centroid method. Seven primary factors were identifiable upon rotation. Several deductions are made relative to the interpretation of the factors and relative to the consistency of the data with the hypotheses which were to be tested. (Courtesy *Psychometrika*)

Coombs, Clyde H. "A Criterion for the Number of Factors in a Table of Intercorrelations" *Psychometrika*, V (1940), 315 (Abstract of a paper read at the September, 1940, meeting of the American Psychological Association)

Cureton, E. E. "Testing in College Personnel Service" *Journal of Consulting Psychology*, IV (1940), 221-24

A survey of the purposes of a college personnel service and of the extent to which available tests lend themselves to such purposes. The need for an adequately standardized differential intelligence battery is emphasized. To show progress and to give information concerning the pattern of abilities and attainments, test scores should be directly comparable. Suggestions are made on coordinating test production. *W. A. Varvel*

Dwyer, P. S. "The Solution of Simultaneous Equations" *Psychometrika*, VI (1941), 101-29

This paper is an attempt to integrate the various methods which have been developed for the numerical solution of

MEASUREMENT ABSTRACTS

simultaneous linear equations. It is demonstrated that many of the common methods, including the Doolittle method, are variations of the method of "single division." The most useful variation of this method, in case symmetry is present, appears to be the Abbreviated Doolittle method. The method of multiplication and subtraction likewise can be abbreviated in various ways of which the most satisfactory form appears to be the new Compact method. These methods are then applied to such problems as the solution of related equations, the solution of groups of equations, and the evaluation of the inverse of a matrix. (Courtesy *Psychometrika*)

Dwyer, P. S. "The Evaluation of Determinants." *Psychometrika*, VI (1941), 191-204

The numerical evaluation of determinants with a modern computing machine is discussed. Various methods are presented and their relations to each other are indicated. The methods presented parallel those developed in the previous papers on "The Solution of Simultaneous Equations." Especially emphasized are the Abbreviated Doolittle and the Compact methods. Additional topics include the evaluation of partially symmetric determinants by means of symmetric methods and the evaluation of determinantal ratios. (Courtesy *Psychometrika*)

Guilford, J. P. "The Difficulty of a Test and Its Factor Composition." *Psychometrika*, VI (1941), 66-77

A factor analysis of the 10 sub-tests of the Seashore test of pitch discrimination revealed that more than one ability is involved. One factor, which accounted for the greater share of the variances, had loadings that decreased systematically with increasing difficulty. A second factor had strongest loadings among the more difficult items, particularly those with frequency differences of two to five cycles per second. A third had strongest loadings at differences of five to 12 cycles per second. No explanation for the three factors is apparent, but the hypothesis is accepted that they represent distinct abilities.

In tests so homogeneous as to content and form, where a single common factor might well have been expected, the appearance of additional common factors emphasizes the importance of considering the difficulty level of test items, both in the attempt to interpret new factors and in the practice of testing. The same kind of item may measure different abilities accordingly as it is easy or difficult for the individuals to whom it is applied. (Courtesy *Psychometrika*)

Guilford, J. P. "A Note on the Discovery of a G Factor by Means of Thurstone's Centroid Method of Analysis" *Psychometrika*, VI (1941), 205-8

A fictitious factor matrix including 16 tests and three factors, one of which was a *g* factor, was prescribed. From it two typical factor problems, including errors of sampling, were derived. Students in training, without awareness of the factor patterns, arrived at essentially correct solutions by the use of Thurstone's centroid method with rotation of axes. Errors in the calculated factor matrix were very close in size to the sampling errors in the correlation coefficients. It is concluded that a *g* factor need not escape detection by Thurstone's procedures if the criteria of complete simple structure are not demanded. (Courtesy *Psychometrika*)

Horst, Paul. "A Non-graphical Method for Transforming an Arbitrary Factor Matrix into a Simple Structure Factor Matrix" *Psychometrika*, VI (1941), 79-99

The most commonly used method of factoring a matrix of intercorrelations is the centroid method developed by L. L. Thurstone. It is, however, necessary to transform the centroid matrix of factor loadings into a simple structure matrix in order to facilitate the interpretation of the factor loadings. Current methods for effecting this transformation are chiefly graphical and require considerable experience and personal judgment. This paper presents a new method for transforming an arbitrary factor matrix into a simple structure matrix by methods almost completely objective. The theory under-

MEASUREMENT ABSTRACTS

lying the method is developed and approximation procedures are derived. The method is applied to a matrix of factor loadings previously analyzed by Thurstone. (Courtesy *Psychometrika*)

Hoyt, Cyril. "Test Reliability Estimated by Analysis of Variance." *Psychometrika*, VI (1941), 153-60

A formula for estimating the reliability of a test, based on the analysis of variance theory, is developed and illustrated. The data needed for the required computation are the number of correct responses to each item and the score for each subject. The results obtained from this formula are identical with those from one of the special cases of the Kuder-Richardson formulation. The relationships of the new procedure to other approaches to the problem are indicated. (Courtesy *Psychometrika*)

Karlin, J. E. "The Isolation of Musical Abilities by Factorial Methods." *Psychometrika*, V (1940), 316. (Abstract of a paper read at the September, 1940, meeting of the American Psychological Association)

Lazarsfeld, Paul F. (Guest Editor) "Radio Research and Applied Psychology." *Journal of Applied Psychology*, XXIV, No. 6 (1940), 661-853

This entire number is devoted to 21 articles dealing with problems of radio research (including two on magazine advertising research techniques), not separately abstracted here because of space limitations. They are classified into the following five groups of papers: I "Commercial Effects of Radio" (F. Stanton, E. Smith & E. Suchman, M. Fleiss, M. Erdélyi), II "Educational and Other Effects of Radio" (S. Reid, J. R. Miles, G. Wiebe), III "Program Research" (J. N. Peterman, H. Schwein, C. Daniel, H. C. Link & P. G. Corby), IV "General Research Techniques" (E. A. Suchman & B. McCandless, M. Rollins, H. Gaudet & E. C. Wilson, D. B. Lucas, R. Franzen, P. F. Lazarsfeld), and V "Measurement Problems" (P. F. Lazarsfeld & W. S. Robinson, C. Daniel, W. S. Robinson, R. Franzen). *F. A. Kingsbury*.

Lentz, Theodore F, and Whitmer, Edith F "Item Synonymization. A Method for Determining the Total Meaning of Pencil-Paper Reactions," *Psychometrika*, VI (1941), 131-9

Items have been studied heretofore for their value as elements of particular tests to the neglect of more fundamental research into the multiple potentiality of items. This article proposes a method of grouping items into "synonymies" comprising all of the items which correlate with a given key item. These synonymies can be used for interpretation of the total meaning of the key item (1) by inspection of the constituent items and (2) by correlational study of obtained single scores of individual persons. The method is illustrated by four items with inter- and intra-correlations, and characteristics of an ideal background reservoir of items are pointed out. (Courtesy *Psychometrika*)

Martin, D. R. "Mental Tests in Clinical Practice" *Australasian Journal of Psychology and Philosophy*, XVIII (1940), 144-53

The author discusses the purpose of mental testing in child-guidance work and describes the battery of tests for general intelligence, special abilities and disabilities, school achievement, personality traits, and emotional stability in use at his clinic. References are made to the recent controversy between Cattell and Vernon over the value of the Binet test. *W. A. Varvel*

McNemar, Quinn "On the Sampling Errors of Factor Loadings" *Psychometrika*, VI (1941), 141-52

The results of three empirical studies on the sampling fluctuation of centroid factor loadings are reported. The first study is based on data which happened to be available on 8 variables for 700 cases and which were factored to three factors for subsamples. The second study is based on fictitious data for 2500 cases which provided separate analyses on 25 samples for each of three situations: five variables, one factor, five variables, two factors, and six variables, three factors.

MEASUREMENT ABSTRACTS

The third study, based on real data for nine variables and 7000 cases, involves separate factorization for 25 samples of 200 cases. The three studies agree in showing that the sampling behavior of first centroid factor loadings is much like that of correlation coefficients, whereas the sampling fluctuations for loadings beyond the first are disturbingly large (Courtesy *Psychometrika*)

McNemar, Quinn "More on the Iowa I Q Studies" *Journal of Psychology*, X (1940), 237-40

In a reply to the Wellman-Skeels-Skodak review [*Psychological Bulletin*, XXXVII (1940), 93-111] of his original critique of the Iowa studies on environmentally-determined changes in I Q, the author does not find it necessary to modify materially his previous criticisms *W A Varvel*

McNemar, Quinn "On the Number of Factors" *Psychometrika*, V (1940), 315 (Abstract of a paper read at the September, 1940, meeting of the American Psychological Association)

Porter, E K "Criteria of a Good Examination" *Public Health Nursing*, XXXII (1940), 558-64

Steps in the construction of a test are outlined. Principles for the construction of tests and test items are presented

Schaefer, Willis C "The Relation of Test Difficulty and Factorial Composition Determined from Individual and Group Forms of Primary Mental Abilities Tests." *Psychometrika*, V (1940), 316-17 (Abstract of a paper read at the September, 1940, meeting of the American Psychological Association)

Van Steenberg, N J "Analysis of Mental Growth of School Children" *Psychometrika*, V (1940), 314 (Abstract of a paper read at the September, 1940, meeting of the American Psychological Association)

MEASUREMENT NEWS*

Dr. John C. Flanagan has been granted a year's leave of absence from the Cooperative Test Service in order that he may accept a commission as a reserve officer in the Army Air Corps. Dr. Flanagan will direct developmental researches and make practical applications with regard to problems of selection of Air Corps personnel.

The authors of the Chicago Reading Tests, Drs. Max D. Engelhart and Thelma Gwinn Thurstone, are conducting an investigation of the comparability of the norms of these tests from form to form and also their comparability with the norms of the Metropolitan and Stanford Reading Tests. Each series of forms of the Chicago tests was standardized independently in successive years by administration to pupils in a representative sample of 30 Chicago elementary schools. Approximately 8,000 elementary pupils took each form when it was given for standardization. In addition, two forms of the sixth-, seventh-, and eighth-grade test were administered in successive years to approximately 8,000 Chicago high-school pupils. The assumption which was made and is now being tested was that norms based on large

*Notes for this department should be sent to Dr. M. W. Richardson, United States Civil Service Commission, Washington, D. C.

samples of pupils drawn from the same schools should be comparable. In the current study each of the three forms of each of the four Chicago tests, and the appropriate Metropolitan and Stanford tests, were administered to several hundred pupils in randomized order to control practice effect. For example, the three forms of Chicago Reading Test D and the Metropolitan and Stanford Advanced Reading Tests were administered to the same elementary pupils. It is planned to determine the equivalence of the raw scores and, on the basis of these data, to make any necessary adjustments to secure precise comparability of the norms.

Professor Karl J. Holzinger of the University of Chicago has written a treatise on Factor Analysis with the assistance of Harry H. Harman, Research Associate. This volume is being published by the University of Chicago Press and will be ready this fall. Professor Holzinger is also joint author of two new monographs on the application of factorial methods. The collaborating authors are M. A. Wenger, Frances Swineford and Harry H. Harman. These monographs will also be published by the University of Chicago Press.

Wright Junior College in Chicago has recently begun a three-year study on the evaluation of terminal education. The study is a part of a comprehensive investigation of junior college terminal education being carried on in nine selected junior colleges by the American Association of Junior Colleges with a grant made by the General Education Board.

The Wright study will attempt to evaluate the present terminal general and terminal occupational programs

MEASUREMENT NEWS

offered at that institution. One of the purposes of the study will be the development of techniques of evaluation for use by other schools.

An extensive measurement program will be initiated in September for the incoming freshman class. Measurements will be made in twelve areas: effective thinking, command of skills and understandings in the major cultural areas, functional understanding of the basic facts of health and disease, interests, appreciations, consumer competence, occupational efficiency, personal-social adaptability, socio-civic consciousness, attitudes, worthy use of leisure, functional philosophy of life. Several measuring devices now available will be used as well as others which are now being developed as a part of the study. Those in the former category include the Cooperative Test Service General Culture and Contemporary Affairs tests, the Kuder Preference Record, and two of the tests of the Progressive Education Association on interpretation of data and nature of proof.

The study is being conducted by a group composed of Dean William H. Conley, Bernard Gold, Alice Griffin, Max D. Engelhart, and Leland Medsker.

Soon to be published by the Social Science Research Council is a monograph on the prediction of personal adjustment. The text has been prepared by Dr. Paul Hoist of Procter and Gamble. The monograph will deal with personal adjustment in connection with vocations, schools, marriage, and criminal recidivism.

EDUCATIONAL AND PSYCHOLOGICAL
MEASUREMENT

Volume I

OCTOBER, 1941

Number 4

CUMULATIVE TEST RECORDS THEIR NATURE AND USES <i>Arthur E. Traxler</i>	323
AN ANALYTICAL DESCRIPTION OF STUDENT COUNSELING <i>E. G. Williamson and E. S. Bordin</i>	341
A COMPOSITION TEST FOR FOREIGN LANGUAGES <i>Lawrence Andrews</i>	355
PERFORMANCE TESTING IN PUBLIC PERSONNEL SELECTION, PART II <i>Sidney W. Koran</i>	365
THE VALUE OF INTELLIGENCE QUOTIENTS OBTAINED IN SECOND- ARY SCHOOL FOR PREDICTING COLLEGE SCHOLARSHIP <i>L. D. Hartson and A. J. Sprow</i>	387
THE THURSTONE MENTAL ABILITY'S TESTS AND COLLEGE MARKS <i>Mary Lou Ellison and Harold A. Edgerton</i>	399
A SHORT CUT IN THE ESTIMATION OF SPLIT-HALVES COEFFI- CIENTS <i>Charles I. Mosier</i>	407
MEASUREMENT ABSTRACTS	409
INDEX FOR VOLUME I	iii

Copyright, 1941, by
SCIENCE RESEARCH ASSOCIATES

STATEMENT OF THE OWNERSHIP, MANAGEMENT, CIRCULATION ETC. REQUIRED BY THE
ACTS OF CONGRESS OF AUGUST 24 1912 AND MARCH 3, 1933
OF EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
Published Quarterly at Chicago Ill. for October 1, 1941

State of Illinois } ss
County of Cook }

Before me a Notary Public in and for the State and county aforesaid personally appeared F. O. Jensen who having been duly sworn according to law deposes and says that he is the Business Manager of the Educational and Psychological Measurement and that the following is to the best of his knowledge and belief a true statement of the ownership, management (and if a daily paper the circulation) etc. of the aforesaid publication for the date shown in the above caption required by the Act of August 24 1912 as amended by the Act of March 3 1933 embodied in section 637 Postal Laws and Regulations printed on the reverse of this form to wit:

1 That the names and addresses of the publisher, editor, managing editor and business managers are: Publisher: Science Research Associates, 1700 Prairie Avenue, Chicago; Editor: O. Frederic Kuder, 1700 Prairie Avenue, Chicago; Managing Editor: Louis H. Lissman, 1700 Prairie Avenue, Chicago; Business Manager: F. O. Jensen, 1700 Prairie Avenue, Chicago.

2 That the owner is: (If owned by a corporation its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding one per cent or more of total amount of stock. If not owned by a corporation the names and addresses of the individual owners must be given. If owned by a firm, company or other unincorporated concern its name and address as well as those of each individual member must be given.) Ralph A. Bard, 208 S. LaSalle St., Chicago, Ill.; Charles S. Boyd, Appleton Coated Paper Co., Appleton, Wis.; R. W. Ginsner, 6400 W. 95th St., Chicago, Ill.; Alfred I. Hamill, 208 S. LaSalle St., Chicago, Ill.; Robert C. McNamara, 629 S. Wabash Ave., Chicago, Ill.; John I. Shaw, 135 S. LaSalle St., Chicago, Ill.; Lylo M. Spencer, 1700 Prairie Ave., Chicago, Ill.; Mrs. Dorothy Bard, c/o Roy L. Bard, 131 S. LaSalle St., Chicago, Ill.; Roy L. Bard, 131 S. LaSalle St., Chicago, Ill.; George M. Bard, II, c/o Ralph A. Bard, 208 S. LaSalle St., Chicago, Ill.; Miss Janet Bard, c/o Ralph A. Bard, 208 S. LaSalle St., Chicago, Ill.; Robert K. Burns, 1700 Prairie Avenue, Chicago, Ill.; Miss Grace M. Wagner, c/o Richard Wagner, 135 S. LaSalle St., Chicago, Ill.; W. C. Winkel, c/o Medina Mfg. Company, Madison, Wis.

3 That the known bondholders, mortgagees and other security holders owning or holding 1 per cent or more of the total amount of bonds, mortgages or other securities are: (If there are none so state.) None.

4 That the two paragraphs next above giving the names of the owners, stockholders and security holders if any contain not only the list of stockholders and security holders as they appear upon the books of the company but also in cases where the stockholder or security holder appears upon the books of the company as trustee or in any other fiduciary relation the name of the person or corporation for whom such trustee is acting is given; also that the said two paragraphs contain statements embracing affiant's full knowledge and belief as to the circumstances and conditions under which stockholders and security holders who do not appear upon the books of the company as trustees hold stock and securities in a capacity other than that of a bona fide owner; and this affiant has no reason to believe that any other person, association or corporation has any interest direct or indirect in the said stock, bonds or other securities than as so stated by him.

5 That the average number of copies of each issue of this publication sold or distributed through the mails or otherwise to paid subscribers during the twelve months preceding the date shown above is: (Not a daily publication.) (This information is required from daily publications only.)

F. O. JENSEN Business Manager

Sworn to and subscribed before me this 2nd day of October 1941

GERTRUDE A. PAYNE, Notary Public

(SEAL)

(My commission expires September 24 1945)

CUMULATIVE TEST RECORDS THEIR NATURE AND USES

ARTHUR E. TRAXLER
Educational Records Bureau

MOST SCHOOLS now recognize certain values in objective tests of academic aptitude and achievement and employ such tests to some extent in the appraisal, placement, instruction, and guidance of their pupils. It is generally understood that objective tests do not measure all educational outcomes, but studies have repeatedly demonstrated that they do measure certain aspects of ability and achievement that are important in the scholastic success of individual boys and girls.

It may seem merely a reiteration of an obvious point to say that the value of a testing program is directly proportional to the nature and extent of the uses of the results by the faculty and students. Nevertheless, emphasis on this point is necessary, for not infrequently school authorities administer a series of tests, file the scores, and then give no further attention to the test data. When no improvement is noted, they blame the tests when the real fault lies with their failure to study the results. Tests are not in themselves remedial instruments, they are tools which can be indispensable aids to diagnosis and thus form an important basis for the planning of instruction and guidance, provided someone carefully and intelligently studies the data which they provide. The analysis of the test results may to some extent be concerned with groups, but it should deal primarily with individuals.

Before test results can be studied and used to best advantage they must be recorded in some convenient form. Alphabetical class lists of the scores and percentile ranks of individual pupils, accompanied by sheets showing the distributions

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

of scores for various classes, are very useful to teachers for purposes of a quick survey of the results for individuals or groups with reference to restricted areas of ability or achievement. They do not, however, readily provide a comprehensive picture of the results of all tests taken by any one individual.

Individual record sheets on which the results from a single testing program are summarized in tabular or graphic form provide a very helpful picture of the status of a pupil at any one time. One can, for example, administer a general achievement test battery such as the Metropolitan or the Stanford, plot the achievement profile of each pupil, and thus obtain a graphic representation of strengths and weaknesses that greatly simplifies the problem of diagnosis. A graph of this type is illustrated in Figure 1. One can see at a glance that in comparison with the grade norms, this pupil is strong in reading, literature, history and civics, and geography, but relatively weak in arithmetic and spelling.

However, distributions, class lists, and diagnostic profiles resulting from one testing program share a common limitation. They show status, but *they do not show growth*. For both instruction and guidance, the concept of growth—how far a pupil has come within a certain period and how far he should be able to go—is probably fully as important as the concept of present status.

Now, there is a type of record that provides evidence about both status at any testing period and growth between testing periods. This is the *individual cumulative record*. It is unquestionably the most valuable aid to the intelligent and efficient use of test results yet devised. It is to other kinds of records what a motion picture is to a snap shot.

The cumulative record presupposes a regular, systematic testing program. If tests are administered in a school at irregular intervals and without definite plan, the value of a record of this kind will be greatly curtailed, but even under these conditions, it will probably prove more useful than any other kind of record of test results.

CUMULATIVE TEST RECORDS

The headings at the top of the chart refer to the tests in the Intermediate Advanced Batteries. The headings at the bottom of the chart refer to the tests in the Primary I and Primary II Batteries as indicated. In order to use the same column for the same subject for all the batteries the first column is used only for Test 1 of the Primary I and Primary II Batteries. The total score column is used only for the Intermediate and Advanced Batteries—Complete. The user of the tests should bear in mind that Test 1 of the Primary II Battery is included in order to get the children interested in the battery and to select those poorest in reading. It is therefore not sufficiently difficult to measure achievement for the third grade pupils or even the better second grade pupils. All Test 1 scores of 15 or 16 were omitted from the total (average) achievement composite independent school medians for indicated grades.

— Pupil's Profile *Values above Grade 90 and below 40 are extrapolated
 --- Median Profile for Grade 7 in Independent Schools

Figure 1

Explanation and Illustration of Cumulative Test Records

Any record for individual pupils that provides for successive additions of the same type of data over a period of years and thus makes possible a study of the changes that have taken place may be called a cumulative record. Thus, a tabular arrangement of test scores and percentiles by years is cumulative in nature. Test results entered in this way, however, cannot readily be apprehended quickly, but require detailed study. This fact has led many persons to favor the graphic representation of test results. Among the various types of graphs thus far devised, the gridiron percentile graph first employed in the American Council cumulative record forms stands out as the most widely used type. Its success is no doubt due partly to the fact that it will accommodate any kind of test result that can be expressed in terms of percentiles.

One of the best known adaptations of the American Council form is the Educational Records Bureau cumulative record for independent schools. This form, like the American Council form, is planned for six years, but it may be expanded to include any number of years. The test portion of this type of record is illustrated in Figures 2, 3, and 4. Let us note in some detail the nature of the information provided.

The card is divided by heavy vertical lines into broad columns, each of which represents a grade or a year in the life of the pupil. The year and grade are indicated at the top of each column.

The front of the card is devoted almost entirely to a record of class work and to an extensive test record. Since the main purpose of this article is to discuss test records, the portion of the sample forms dealing with subjects, marks, and credits has not been filled out. The test results are reported in both tabular and graphic form. The scores and corresponding percentiles are entered in the table and the percentiles are then used as the basis of the graph, which occupies approximately the lower half of the card.

The graph of test scores is the clearest phase of the record to one familiar with graphs of this kind, but it often seems

CUMULATIVE TEST RECORDS

[illegible]

somewhat puzzling to persons who have had no experience with it. The percentiles along the scale at the left are placed according to standard deviations in a normal distribution, and thus the distance between successive percentiles is much smaller near the median than near the extremes. The median, or 50th percentile, is marked by the heavy line going horizontally across the graph. The symbols at the top—Jy, Au, S, O, and so forth—stand for the months of the year. The months are grouped according to the school year rather than the calendar year.

The same percentile data that are shown in the table of scores are entered in the graph, except that to prevent overcrowding, the percentiles for the parts of the English test have been omitted from the graph. The percentiles used in these records are based on results in independent schools, but the interpretation of public-school percentile ratings would be made in exactly the same way.

The small dots on the graph show the placement of the various percentiles, the dots being identified by the abbreviated names of the tests printed near them. For example, in Figure 2, the dot toward the top of the graph is labeled "French" to indicate that it stands for the percentile on the Cooperative French test. The percentile for the pupil's total score of 61 on the French test is 93, and this is indicated by placing the dot opposite 93, one of the points shown on the percentile scale at the left of the chart. In other words, the pupil's French score was above the scores of 93 per cent of the independent-school ninth-grade first-year French students who took the test in the spring of 1941.

The percentile points for tests that are in the same field from year to year are connected by lines, so that one can readily follow a particular type of achievement throughout the whole period covered by the test. For instance, one of the lines in Figure 3 runs from the arithmetic percentile in Grade 6 to the arithmetic percentile in Grade 7, and from that point to the arithmetic percentile in Grade 8, thence to the elementary algebra percentile in Grade 9, et cetera. Achieve-

CUMULATIVE TEST RECORDS

ment percentiles are connected with solid lines, academic aptitude percentiles with broken lines, and chronological age percentiles with dotted lines

The record shown in Figure 2, that of Edwin Martin,¹ covers only one year. In many independent schools, a considerable proportion of the records will be of this type, for the number of one-year students attending private schools tends to be fairly large. Even in the case of single-year records, it is desirable to record the data on a cumulative record form, for such a procedure facilitates comparisons between academic aptitude and achievement and makes it possible to summarize readily the student's test record for the year as a whole.

Figure 2 shows that in the fall of 1940, Edwin was close to, but slightly below average for his grade in chronological age and that he was somewhat below the median in academic aptitude, as indicated by the results of the American Council Psychological Examination, and in reading, as measured by the Nelson-Denny Test. These results are recorded directly below the letter O, which shows that the data were obtained in October.

The spring, 1941, percentiles are entered beneath the letter A, and thus one knows that the tests were given in April. Edwin seems to be an able student of foreign language. As already indicated, his French score was above those of all but 7 per cent of the first-year French students in Grade 9. His total score on the Latin test fell within the highest third of the independent-school ninth-grade first-year group.

In science and elementary algebra, the boy was above the independent-school median but not outstanding. His total English percentile and his literary acquaintance percentile were below the median but above his academic aptitude percentile.

In general, Edwin's achievement test percentiles were somewhat higher than his percentiles in academic aptitude and reading. This is, of course, an encouraging finding, for it indicates that, presumably because of application and hard

¹These are actual test records, but the names of the pupils and the schools are fictitious.

work, the boy's achievement record near the end of the school year was better than one would expect it to be on the basis of the fall test results

Let us now examine a cumulative record covering several years. The test record of Betty A. Stetson, as shown in Figure 3, is that of a girl who was a little younger than the average pupil in her grade but who was generally high in both academic aptitude and achievement. Her Otis intelligence test percentiles in Grade 6 were exceptionally high. Her later academic aptitude percentiles were a little lower, but all of them were significantly above the median for her grade. In fact, her Otis scores in Grades 6 and 7 and her scores on the American Council Psychological Examination in Grades 9 and 10 were in the highest tenth of the scores made by the independent-school pupils at the same grade levels.

Betty's achievement test percentiles were, in general, somewhat lower than her percentiles in academic aptitude, but most of them were in the upper half of the scores of the pupils in her grade. The only achievement percentiles below the median were those for spelling in Grade 6, geography in Grade 7, arithmetic in Grade 8, general science in Grade 10, and modern European history in Grade 11. The history score was the only very low result in the entire record.

This girl is obviously an excellent reader. On the reading tests, she maintained a position within the highest tenth of her grade throughout the entire period.

In a graphic record of this kind, growth in any subject precisely equal to the growth of the group as a whole in that subject is shown by an exactly horizontal line. That is, if a pupil improves just as much as the group improves in a year, he will maintain the same percentile rating from one year to the next. Lines which go upward, then, indicate greater than average growth and lines which slope downward suggest less than average growth. In interpreting such variations, however, one should keep in mind the fact that every test involves a certain amount of sampling error and that the population on which the percentiles are based is not exactly the same from

CUMULATIVE TEST RECORDS

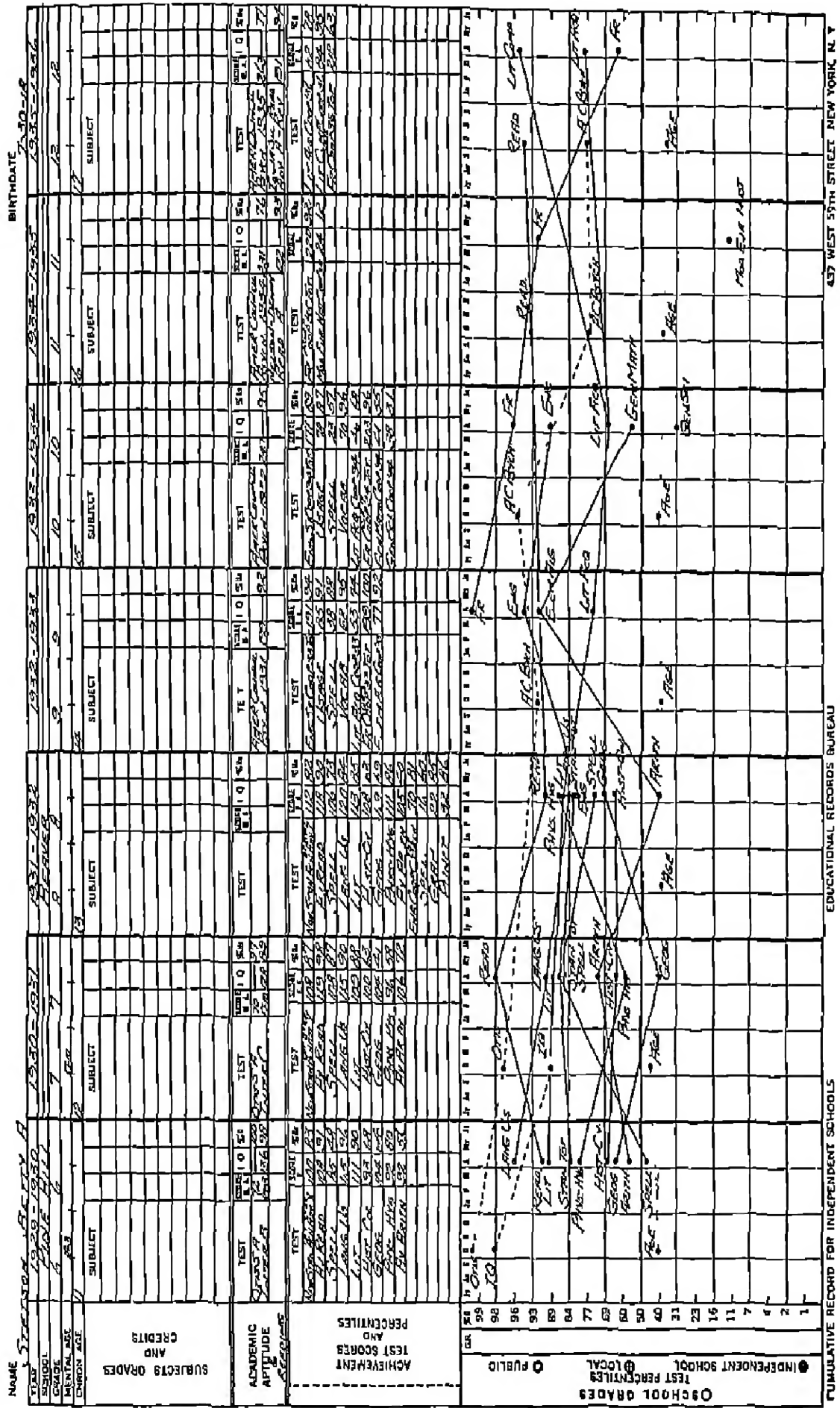


Figure 3

year to year. So some variation in percentiles for the same subject is to be expected even when growth is normal. One should also remember that tests of different subjects in the same field—for example, algebra and plane geometry or biology and chemistry—involve somewhat different abilities. In these instances, changes in percentile rating should not be interpreted in terms of growth.

However, marked gains or losses in percentiles on the same test, such as an English test or a foreign language test, are symptomatic, and they may be indicative of a need for counseling attention in order to find the reason for the variation. For example, the steady downward trend of Betty's percentiles in French would seem to require an explanation that cannot be obtained from the record itself.

On the whole, the seven-year record shown in Figure 3 indicates definitely superior ability and achievement. A counselor, school principal, or college admissions officer familiar with this type of record could decide almost at a glance that, as far as aptitude and attainment are concerned, this girl should do well even in a highly selective college.

The record of Charles W. Loring, shown in Figure 4, also includes Grades 6 to 12, inclusive, but it covers a period of eight years rather than seven, since this boy repeated the eighth grade. The general level of the percentiles is in marked contrast to that of the percentiles on the preceding record. This is an over-age pupil who is rather low in both academic aptitude and achievement in comparison with the average for his grade. Because he is advanced in chronological age, the percentiles corresponding to mental age and raw scores on intelligence tests tend to be somewhat higher than the percentiles for I.Q., but with one exception they are below the independent-school median.

In Grade 6, all but one of Charles' scores on the New Stanford Achievement Test were distributed below the median for independent-school pupils at that grade level. The next year, most of his achievement test percentiles went upward to some extent, but the percentiles for language usage and arith-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

metic were the only ones above the independent-school median for Grade 7. In the following years, nearly all his percentiles were below the median for his grade, although some of them were not far below it.

There is some evidence that this boy had read rather widely and that he was a fairly competent reader. Three of his literature percentiles and three of his reading percentiles were above the median for his grade.

Charles' repetition of Grade 8 did not significantly raise his subsequent record on the achievement tests. The only line that went upward noticeably as a result of the repetition of this grade was the one for chronological age. The whole record is that of a pupil who probably should not attempt to enter the usual liberal arts college after graduation from the secondary school. Rather, he needs guidance into preparation for some type of vocation the demands of which are consistent with his mediocre scholastic attainments.

It will be observed that there is a general consistency in the test results throughout both illustrative records. The girl whose record is shown in Figure 3 was high in academic aptitude and achievement in the elementary school and she maintained this superiority throughout the secondary school. The boy whose record is contained in Figure 4 was low in academic aptitude and achievement in the elementary school and this low record was continued in the secondary school. In both cases, the general level of the percentile ratings in Grade 12 could have been predicted from the results of the achievement tests taken in Grade 6.

The tendency of the cumulative record of test results for an individual pupil to be in agreement from year to year is one of the most noteworthy aspects of this type of record. This tendency is verified by hundreds of such records which have been prepared at the Educational Records Bureau and other institutions. While the percentiles on an occasional test may vary markedly in successive years, the whole picture of a pupil's record tends to remain much the same. This is

CUMULATIVE TEST RECORDS

usually true regardless of transfer from one school to another or variation in type of instruction. Comprehensive test records obtained even as low as the second or third grade often predict with remarkable fidelity the level of achievement a pupil will attain in his senior year of high school. The fact that test results distributed over a long period of time tend to be positively correlated causes cumulative test histories to have exceptional potential guidance values.

Notwithstanding the general tendency just indicated, it is true that a pupil's test record for an entire year may sometimes be decidedly out of line with his scores in other years. When this happens, an explanation that can be made only in the light of much other information about the pupil is required. Consequently, a cumulative record of other kinds of data is needed if one is to make an adequate interpretation of the test record. For this reason and many other reasons, it is advisable for schools to maintain cumulative records that cover not only test results but that include home background, class work, interests and activities, personality adjustment, and various other factors. The interrelationships of the different kinds of information that can be recorded on a form similar to the American Council card are brought out by the record for Harry Connelly, as shown in Figure 5.

It is evident from Figure 5 that this boy tended to be below the independent-school median in his scores on the Metropolitan Achievement Test taken in Grade 8, but that he was consistently above the median in scores on all the achievement tests taken in Grades 9 to 12. He was especially high in English, literary acquaintance, and science. The boy's superior test record in the four high-school grades agrees with his consistently high percentile ratings on the academic aptitude tests.

It appears that an explanation of the marked difference between Harry's test scores in Grade 8 and his test scores in the later years is to be found in the data entered on the back of the card. Although he was obviously bright, he was lazy and disorderly in the eighth grade and it is probable

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

NAME
CONNELLY Harry

BIRTHDATE
1-5-21

YEAR	SCHOOL GRADE	MENTAL AGE	CHRON AGE	1925-1926		1926-1927		1927-1928	
				SUBJECT	SCORE	SUBJECT	SCORE	SUBJECT	SCORE
SUBJECTS GRADES CREDITS AND	ENG 1	64	71	ENG 10	82	ENG 11	88	ENG 12	90
	FR 1	62	70	FR 10	71	FR 11	80	FR 12	85
	Lat 1	65	72	Lat 10	77	Lat 11	85	Lat 12	88
	COMPLAT 1	62	70	COMPLAT 10	77	COMPLAT 11	85	COMPLAT 12	88
	GEN SCIENCE 1	64	71	GEN SCIENCE 10	77	GEN SCIENCE 11	85	GEN SCIENCE 12	88
	HIST 1	71	77	HIST 10	80	HIST 11	88	HIST 12	90
	PHYSICS 1	64	71	PHYSICS 10	77	PHYSICS 11	85	PHYSICS 12	88
	CHEM 1	64	71	CHEM 10	77	CHEM 11	85	CHEM 12	88
	PHIL 1	64	71	PHIL 10	77	PHIL 11	85	PHIL 12	88
	ART 1	64	71	ART 10	77	ART 11	85	ART 12	88
ACADEMIC ATTITUDE AND READING	ENG 1	64	71	ENG 10	82	ENG 11	88	ENG 12	90
	FR 1	62	70	FR 10	71	FR 11	80	FR 12	85
	Lat 1	65	72	Lat 10	77	Lat 11	85	Lat 12	88
	COMPLAT 1	62	70	COMPLAT 10	77	COMPLAT 11	85	COMPLAT 12	88
	GEN SCIENCE 1	64	71	GEN SCIENCE 10	77	GEN SCIENCE 11	85	GEN SCIENCE 12	88
	HIST 1	71	77	HIST 10	80	HIST 11	88	HIST 12	90
	PHYSICS 1	64	71	PHYSICS 10	77	PHYSICS 11	85	PHYSICS 12	88
	CHEM 1	64	71	CHEM 10	77	CHEM 11	85	CHEM 12	88
	PHIL 1	64	71	PHIL 10	77	PHIL 11	85	PHIL 12	88
	ART 1	64	71	ART 10	77	ART 11	85	ART 12	88
ACHIEVEMENT TEST SCORES AND PERCENTILES	ENG 1	64	71	ENG 10	82	ENG 11	88	ENG 12	90
	FR 1	62	70	FR 10	71	FR 11	80	FR 12	85
	Lat 1	65	72	Lat 10	77	Lat 11	85	Lat 12	88
	COMPLAT 1	62	70	COMPLAT 10	77	COMPLAT 11	85	COMPLAT 12	88
	GEN SCIENCE 1	64	71	GEN SCIENCE 10	77	GEN SCIENCE 11	85	GEN SCIENCE 12	88
	HIST 1	71	77	HIST 10	80	HIST 11	88	HIST 12	90
	PHYSICS 1	64	71	PHYSICS 10	77	PHYSICS 11	85	PHYSICS 12	88
	CHEM 1	64	71	CHEM 10	77	CHEM 11	85	CHEM 12	88
	PHIL 1	64	71	PHIL 10	77	PHIL 11	85	PHIL 12	88
	ART 1	64	71	ART 10	77	ART 11	85	ART 12	88
SCHOOL GRADES TEST PERCENTILES LOCAL INDEPENDENT SCHOOL PUBLIC	ENG 1	64	71	ENG 10	82	ENG 11	88	ENG 12	90
	FR 1	62	70	FR 10	71	FR 11	80	FR 12	85
	Lat 1	65	72	Lat 10	77	Lat 11	85	Lat 12	88
	COMPLAT 1	62	70	COMPLAT 10	77	COMPLAT 11	85	COMPLAT 12	88
	GEN SCIENCE 1	64	71	GEN SCIENCE 10	77	GEN SCIENCE 11	85	GEN SCIENCE 12	88
	HIST 1	71	77	HIST 10	80	HIST 11	88	HIST 12	90
	PHYSICS 1	64	71	PHYSICS 10	77	PHYSICS 11	85	PHYSICS 12	88
	CHEM 1	64	71	CHEM 10	77	CHEM 11	85	CHEM 12	88
	PHIL 1	64	71	PHIL 10	77	PHIL 11	85	PHIL 12	88
	ART 1	64	71	ART 10	77	ART 11	85	ART 12	88

CUMULATIVE TEST RECORDS

that he lacked both the preparation and the interest in the test itself that would be required for high scores on the Metropolitan test, or any other test of general achievement. There was much improvement in the boy's behavior and attitude in the ninth grade and in subsequent grades, and consequently his achievement increased until it was proportionate to his ability. The whole picture is that of an intelligent, able boy who was immature in behavior and attitude but who became much more mature during the secondary-school years. At the end of the secondary school, he unquestionably had the ability and the preparation for better than average college work. Experience indicates that a record of this kind furnishes a far better basis for prognosis of college success than is provided by a transcript of credits and an admission form filled out by the school when the pupil is near the end of his secondary-school course.

Cumulative records in terms of percentiles are not the only kind of graphic record of test results that can be used. If the results of all tests employed in a school's program are expressed in terms of standard scores, Scaled Scores, or some other comparable unit, the data may be graphed on that basis. Such units are sometimes preferable to percentiles for purposes of showing growth and if they take their origin from a common standardization group, as do the Scaled Scores of the Cooperative Test Service, the influence of the selective factor found in certain subjects, such as the foreign languages, is obviated.²

It should be clearly understood that cumulative records of test results can be kept without the use of any graph whatsoever. The preparation of the graphic part of the record is, of course, a time-consuming clerical job. While it is a distinct aid to interpretation, schools in which the time and cost

²For an illustration of a cumulative record based on Scaled Scores see John C. Flanagan, *The Cooperative Achievement Tests: A Bulletin Reporting the Basic Principles and Procedures Used in the Development of Their System of Scaled Scores*, p. 37. New York: The Cooperative Test Service, December, 1939.

of the graph would be prohibitive can maintain usable test records in tabular form. The main thing is to record the data in organized fashion so that trends can be discerned.

Uses of Cumulative Test Records

As already indicated, the uses that are made of cumulative test records depend largely on the interest, initiative, and understanding of the administration and faculty in each local school. Among the possible uses of such records are the following:

- 1 Counselors may use cumulative test records in conferring with pupils and guiding them toward educational and vocational choices consistent with their ability and achievement.
- 2 Teachers may study them in order to plan their instruction to accord with the aptitude, knowledge, and understanding of the individuals in their classes.
- 3 Administrators and personnel officers may refer to them when conferring with parents about their children.^a
- 4 Principals and guidance directors may take them into consideration when recommending graduates to colleges or to prospective employers.
- 5 College admissions officers may use them as one type of evidence on which decisions about admitting applicants are based. In order to conserve the time of the college admissions officer, the school should of course include a paragraph of interpretation when the record is sent to the college. Admissions officers expect to receive from the school an estimate of a candidate's fitness and they will place more credence in the estimate when it is based in part upon tangible information.
- 6 Schools may employ them in placing transfer pupils in courses to which they are suited.
- 7 Administrative officers and department heads may use them in sectioning classes on the basis of ability.

^aAn excellent discussion of this type of use is given in Robert N. Hilkert, "Parents and Cumulative Records," *Educational Record*, Supplement No. 13, pp. 172-83. Washington, D. C.: American Council on Education, January, 1940.

CUMULATIVE TEST RECORDS

8 Remedial teachers may consult them in selecting pupils for special remedial work and in planning that work

9 Psychologists and psychiatrists may turn to them for leads in diagnosing personality maladjustments and planning treatment

10 Superintendents and principals may make limited use of them in appraising the work of the school and introducing modifications This type of use should be carefully thought out and cautiously applied

11 Counselors and teachers may refer to them as a means of stimulating pupils to do their best work This is a legitimate use if the comparison is directly with the previous record of each pupil and only indirectly, or not at all, with that of other pupils⁴

12 Finally, the entire faculty may employ cumulative test records in developing what is perhaps the school's most important function—the planning for each pupil of a program that is suited to him and the individualization of instruction in accordance with such a program⁵

The American Council cumulative record forms, from which the card used in the illustrations in this article was adapted, are now being revised⁶ A tentative draft of the revised high-school form is ready and it will be tried out soon in several public high schools Changes have been made in various parts of the record to accord with modern trends in education It is significant that in the revised form the test

⁴The use of test records in pupil self-appraisal is described in Richard D. Allen, *Self-Measurement Projects in Group Guidance*, Inor Group Guidance Series, Volume III (New York: Inor Publishing Company, Inc., 1934), xviii+274

⁵See Ben D. Wood, "The Need for Comparable Measurements in Individualizing Education," *Educational Record*, Supplement No. 12, pp. 5-13. Washington, D. C.: American Council on Education, January, 1939.

⁶The revision of the American Council cumulative record forms is being done by a subcommittee of the Committee on Measurement and Guidance of the American Council on Education. The chairman of the subcommittee is Eugene R. Smith, Beaver Country Day School, Chestnut Hill, Massachusetts.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

section continues to be one of the most important aspects of the record. Any forward-looking cumulative record, regardless of whether it is devised by an organization of national scope or by a local school system, will inevitably include a thorough test record, for it is becoming generally recognized that a prerequisite to an adequate program of guidance is a comprehensive, systematic testing program.

AN ANALYTICAL DESCRIPTION OF STUDENT COUNSELING¹

E G WILLIAMSON

and

E S BORDIN

University of Minnesota

THE OBJECTIVES of the student counseling program at the University of Minnesota have not changed fundamentally since its inception in the Arts College in 1923. The aims of the group of Arts College faculty counselors as stated in 1928 by Paterson were "First, to bring about a more harmonious adjustment of individual students to the opportunities available within and without the University, and second, to establish, as far as possible, a friendly and constructive relationship between individual members of the faculty and students desiring such contact" (2 265-266).

Subsequent trends in the University's curricular organization and the developing facilities for personnel work have led to a greater differentiation of function within the total counseling program.² The increasing complexity and the consequent professionalization of certain types of counseling resulted in the establishment of the University Testing Bureau among other specialized agencies for the treatment of student problems. This University-wide counseling agency is both coordinate and coordinated with the counseling agencies of the separate colleges within the University.

¹Assistance in the tabulation and summarization of materials was provided by Minnesota work projects under project 6714, sub-project 85, sponsored by the University of Minnesota.

²For an historical treatment of these developments see E G Williamson and T R Sabin (Minneapolis: Burgess Publishing Company, 1940) 115 pp.

The Testing Bureau, in its function as a counseling agency,³ provides professionalized educational and vocational guidance supplementary to the services of other personnel departments on the campus. Counseling is performed on an individualized basis, the counselor using information from tests, reports from other personnel workers on the campus and from community and high school agencies, and from clinical interviews with the student.

In a series of papers published recently, the authors treated the problem of the evaluation of these counseling services. The first in the series presented a systematic analysis of experimental methods as applied to this type of counseling (4). Based upon our conclusions with regard to method, two experimental evaluations of this counseling were reported.

The first of these experiments (5) investigated the relative adjustment of students who did and did not cooperate with the counselor. The criteria of adjustment and cooperation were judgments by workers who had not been involved in the counseling process and were based on readings of the case history and follow-up interviews. The results showed that students who cooperated were more likely to be adjusted.

The second experiment (6) tested the hypothesis that students counseled by the Bureau would be better adjusted and more successful academically than students who had not been counseled by the Bureau or any college counseling agency. This hypothesis was found to hold for the comparison of a counseled with a matched non-counseled group of freshman Arts College students. The criteria in this study were judgments of adjustment and cooperation and average grade achievement.

Future progress in counseling of this nature will depend upon knowledge of the resources and techniques utilized by the counselors, the types of problems dealt with, and the effectiveness with which these problems were handled. In

³The Bureau also functions as a University-wide testing agency and as a locus for research in testing and counseling.

ANALYTICAL DESCRIPTION OF STUDENT COUNSELING

this paper we are concerned with giving a representative picture of the resources and general techniques utilized in the University Testing Bureau over the period from 1932 to 1935

Analyses of faculty counseling in the Arts College (3, 8) and an exploratory analysis of Testing Bureau counseling (4 253-260) form the background for the present study. The exploratory study of Bureau counseling was based on a sampling of 196 student cases, analyzed as to origin, class, and college. The representativeness of these cases in terms of high school scholarship and college aptitude score was determined. Summaries were presented of the kinds of case data used, the agencies consulted or referred to for diagnosis and treatment, the types and frequencies of student problems, and the general counseling techniques used.

The present study is designed to amplify the description of Bureau counseling from a much broader sampling of the total case load. A total of 2053 student cases, the bulk of students who came in for complete counseling services over the period from 1932 to 1935, formed the population for this survey.¹ The actual case history folders, including records of counseling interviews, were analyzed, and the presence of certain items of information tabulated. No questionnaires filled out by the students were used for this analysis. This study, therefore, provides an answer to the question, "What is counseling?" in terms of the judgments of counselors made in terms of particular students and not of students in general.

Of the 2053 cases, 1223 students were men and 830 were women. Classified according to year in college, there were 617 pre-college students (recent high school graduates), 721 freshmen, 482 sophomores, 143 juniors, 54 seniors and 36 graduate students. By college the distributions were: General College, 428, Arts College, 1038, pre-college who did not matriculate in the University, 197, Chemistry-Engineer-

¹By "complete counseling services" is meant testing and extensive interviewing. The Bureau also provides many types of testing services for members of the student personnel staff of the University.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

ing-Mines, 133, Agriculture, 79, Education, 48, Business, 34; Medical-Dental-Pharmacy, 23, Graduate School, 36, Nursing, 18, and University College, 5

Origin of the Cases

The efficiency of a counseling program must, in part, be measured on the basis of its integration and coordination with other personnel functions. This criterion implies that counselors at various levels of specialization are aware of the limits of their functions and are making use of the services of specialized personnel workers through the medium of referral.

With this in mind the origin of the cases counseled in the Bureau becomes pertinent to its efficiency. Obviously, the two main categories of origin are referred and voluntary. Over fifteen years of experience in the counseling program at Minnesota have shown that the best results can be achieved when the student comes voluntarily to the counselor or seeks assistance at the suggestion, but not command, of some member of the University staff or student body. Of the total of 2053, 1069 of the students were classified as voluntary cases. Actually, of the 984 remaining students classified as referred cases, in only 93 cases was referral made by University officials in the spirit of pressure. These were students of low scholarship who had been referred to the Bureau as part of procedures involved in scholastic discipline. A total of 791 students had been referred to the Bureau for testing and counseling after interviews with a college counselor or faculty member. In addition, 100 students had been referred by high school counselors or community welfare agencies.

The largest proportion of the voluntary cases, 892, came to the Bureau after having heard about its services through bulletins, class lectures, friends, or relatives. In addition, 122 students were told about the Bureau's service in the Registrar's office and 55 had learned about it from high school or college teachers other than counselors.

What distinguishes these students coming by way of various campus and community agencies? Analyses of the vari-

ANALYTICAL DESCRIPTION OF STUDENT COUNSELLING

ance (1 Chap V) in high school percentiles and college aptitude test scores give a partial answer to the question. Referred students tend to be lower than voluntary students in both high school achievement and college aptitude (F values of 24.08 and 76.59, respectively, both well beyond the one per cent point). While there are significant variances between referred and voluntary cases, "t" tests (1.97) revealed that the variation of the sub-categories in college aptitude was homogeneous within each of the two main divisions. This means that there were no reliable differences in ability, either among students who had been referred by a faculty counselor, a college official, or a high school counselor or welfare officer, or among students who came voluntarily as a result of contact with high school or college faculty, the Registrar's office, or some informal source of information about the Bureau's services.

In the case of high school achievement, differences do exist between students referred by high school counselors or welfare agencies and students referred by college officials, college counselors, or faculty members. The relations found between type of problems and origin of case tend to clarify the picture. Students who have been referred by high school counselors or welfare agencies are more likely to have financial and health problems and are less likely to have vocational and educational problems. Thus we may conclude that students referred from these sources outside the University are likely to be students with the financial or health problems referred because they were good students in high school and had well-developed vocational goals.

The relation of type of problem to origin of case also gave indications that the other two types of referred students tend to have fewer financial problems and that students who came voluntarily were more likely to have vocational problems and less likely to have health problems.

Types and Frequencies of Problems

Before more adequate evaluations of counseling can be made, the types and frequencies of problems encountered must

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

be described. Then only can the foundation be laid for more precise experimentation in which treatment methods are differentiated according to their relative value for various types of problems or problem constellations.

The scope of this paper is limited to a general description of counseling. Another paper is being prepared which will describe how these problems cluster for the students counseled and with what characteristics these clusters are associated. This is conceived as the first step toward evolving a symptomatology for use in counseling.

TABLE 1

TYPES OF STUDENT PROBLEMS AS RECORDED IN CASE HISTORIES

	Frequency of Occurrence
A. Financial	
1. Need or desire for part time work, scholarship or loans, inadequate finances	443
B. Vocational	
1. Poor aptitude for chosen vocation	270
2. Inability to decide between two or more vocational choices	756
3. Definite choice but wants confirmation or encouragement	976
4. Definite choice but in doubt about aptitude	116
5. Definite choice but based only on influence of family, friends, etc.	108
6. Dearth of interest in any vocation	42
7. Information needed about occupations in general	66
8. Vocational choice without adequate self-analysis	60
9. Inadequate information in regard to professional choice	50
	2444
C. Educational	
1. Poor aptitude for college work	292
2. Selection of course in line with occupational choice	632
3. Inferiority in academic skills such as reading, study habits, English usage, etc.	446
4. Understanding grading standards	15
5. High general aptitude and poor scholastic achievements	184
6. Understanding responsibilities in college	134

ANALYTICAL DESCRIPTION OF STUDENT COUNSELING

7	High aptitude hampered by standard curricula	20	
8	Outside work interfering with studies	71	
9	University entrance without proper requirements	48	1842
D Social, Personal, and Emotional			
1	Too much social life or too many social activities	43	
2	Inadequate participation in extra-curricular activities	23	
3	Selecting student activities in line with interests	8	
4	Social personality traits which may hinder professional success	115	
5	Need for encouragement and self-confidence	230	
6	Social timidity	89	
7	Emotional disturbances	172	
8	Family domination in vocational choice	78	
9	Conflict with family or friends	91	
10	Parental anxiety for a wise vocational choice	52	
11	Fear of intellectual inadequacy	34	
12	Idealization of a profession	17	
13	Over-evaluation of a college degree	17	969
E Health and Physical Disabilities			
1	Serious physical disabilities	107	
2	Easily fatigued	27	
3	Inability to do justice to work because of intermittent illness	35	
4	Physical habits, diet and sleep, etc	9	178
			<hr/> 5876

Table 1 shows the distribution of the 5876 problems found in the case records of these 2053 cases. We see that about two-thirds of the problems of the students were of an educational or vocational nature. The most frequent vocational problems were found among cases of students unable to decide between two or more vocational choices or who wanted confirmation or encouragement in making a vocational choice. The most typical educational problems were those of selecting a training program appropriate to the vocational choice and those due to inferiority in such academic skills as reading, study habits, and English usage.

Social-personal-emotional problems were the next most frequent types of problems. Modal sub-types were less marked

here. The three most frequent types of social-personal-emotional problems were need for encouragement and self-confidence, social personality traits which may hinder professional success, and emotional disturbances, for the most part of a non-psychiatric nature.

Health problems were infrequently found among these cases, 178 problems being discovered and reported to the Bureau by the University Student Health Service.

The Informational Basis of Counseling

The progress of counseling as a discipline has been characterized by a departure from "gold brick" methods of judging abilities and character. The trend is toward a greater reliance upon the systematic collection of data about the individual by means of standardized tests, reports from other people who have had contacts with him under diverse conditions, medical records, and so on. The only vestige of early counseling methods is the interview. This is still the most vital part of the counseling process, but is now the melting pot in which the student and the counselor integrate information to draw out a unified picture of the individual and to plan the next steps in adjustment.

In Table 2 we present a summary of the number and frequencies with which each source of data was consulted by the counselor. If frequency and source of data are grouped together, 27,866 units of data were used as the basis for counseling. The most frequent source of information was vocational and educational tests given in the Testing Bureau. Clearance slips from the Faculty-Student Contact Desk (6.83) provided information in 2038 cases.

Other important sources of data were Health Service reports, University Entrance Test rating, and grades from the Registrar's office. The fact that reports from family or other relatives were the least frequent sources of information could be taken as an indication of the need for a study to determine whether a social worker would be a valuable addi-

ANALYTICAL DESCRIPTION OF STUDENT COUNSELLING

NUMBER AND SOURCE OF DATA													
Number of Data in each Case	Reports from Health Service	Univ Entrance Test Rating	Grades from Registrar's Office	Vocational and Educ Tests, No Given by U T B	Clear Slip from Faculty-Student Contact Desk	Special Test Results	Application Blanks	Reports from College Deans	Reports from Faculty or Other Relations	Reports from II S Faculty Counselor	Reports from Others	Req of No of Data in Each Case	Total Data Collected
1	1045	109	873	19	728	128	11	222	5	10	45	3195	3195
2	96	415	260	33	786	1		35	2		24	1652	3304
3	4	1253	64	88	372	1		12			8	1802	5406
4	1	158	8	166	111			8			3	455	1820
5		55	1	252	27			1			1	337	1685
6			2	326	10							338	2028
7		1		315	4							320	2240
8			1	281								282	2256
9				163								163	1467
10				116								116	1160
11				62								62	682
12				47								47	564
13				38								38	494
14				28								28	392
15				16								16	240
16				18								18	288
17				8								8	136
18				6								6	108
19				1							1	2	38
20				3								3	60
21				5							1	6	126
22				3								3	66
23				1								1	23
26				1								1	26
30				1								1	30
32				1								1	32
Total of each Type of Data Collected													27,866
	1253	5612	1642	14,574	4083	133	11	365	9	10	174		

tion to the Testing Bureau staff. It would be necessary to determine how much information would be added and how useful that information would be in diagnosis and treatment.

Counseling Procedures Classified by Type of Problem

One of the areas for research in counseling which has been least exploited is the precise description of the counseling interview—the relationships between various counseling processes and the effectiveness with which each type of problem is handled. The ultimate objective of such description is the delineation of these counseling processes in terms of fundamental psychological categories. One possible preliminary step may be the general description of what the counselor did with regard to various types of problems. This is called a general description because it does not attempt to take into account the psychological setting within the interview (e.g., the specific attitudes the counselor and counselee had toward each other at that point) when the behavior described occurred.

In Table 3 we present a general description of the Bureau's counseling procedures. The data show that the commonest procedures in counseling students with financial problems took the form of discussing the need for work, discussing scholarships and loan funds as a source of money, and discussing the relation of part-time work to the student's class schedule.

In the treatment of vocational problems, the counselors relied mainly on discussions of aptitude and on advice and recommendations of occupational choice on the basis of test results. Other frequent procedures include advising vocational "tryouts" through college courses, descriptions of occupations and advising a general background training before a definite choice is made.

With educational problems, the most frequent procedure was aid in selecting a schedule of classes in line with aptitude. In another large number of cases the counselors discussed course prerequisites, sequence of courses, and the like. Attempts at cultivation of interests in studies and scholastic

ANALYTICAL DESCRIPTION OF STUDENT COUNSELING

TABLE 3

COUNSELING PROCEDURES CLASSIFIED BY TYPE OF PROBLEM

Types of Problems		Frequency of Occurrence
<i>Financial Problems</i>		
1	Discussion of relation of part-time work to class schedule	71
2	Discussion of need for work	83
3	Suggestions of ways of getting jobs	12
4	Discussion of student's expenses and financial resources	54
5	Discussion of scholarship and loan funds	78
6	Letters of recommendation for jobs, scholarships and loans	40
7	Referral to employment bureau	22
8	Referral to financial aid agencies	31
		391
<i>Vocational Problems</i>		
1	Description of occupations	232
2	Referral to informational books	128
3	Discussion of aptitude	1120
4	Discussion of student's financial resources for occupational training	125
5	Vocational tryouts through college courses	351
6	Advice and recommendation of occupational choice (on basis of test results)	1346
7	Advice of general background training before definite choice is made	251
8	Discussion of method of entering and securing employment in chosen occupation	120
		3673
<i>Educational Problems</i>		
1	Use of class schedule for program making	4
2	Discussion of course prerequisites, sequence of courses, etc	304
3	Cultivation of interests in studies, scholastic record, etc	121
4	Explanation of recitation method of study	13
5	Discussion of special surroundings conducive to effective study	31
6	Discussion of methods of vocabulary building	27
7	Tutorial aid with specific subjects	20
8	Aid in selecting a schedule of classes in line with aptitude	642
9	Aid in budgeting hours for study	49
10	An attempt to analyze cause for difficulty with a specific subject	38
11	Explanation of student's low aptitude as cause of low scholarship	34
12	Recommendation of non-college type of vocational training	149

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 3 (Continued)

Types of Problems		Frequency of Occurrence
13	Attempt to diagnose and overcome special disability in spelling, grammar, mathematics, etc	25
14	Discussion and explanation of student's responsibility in college, grading standards, etc.	74
15	Recommend student change course of study	182
16	Discussion of eligibility for prescribed work	97
17	Referral to "How to Study" instructors	62 1872
Social, Personal and Emotional Problems		
1	Warning of over-emphasis put upon his social activities	13
2	Arranging contact with proper activities such as band and debate, etc	1
3	Suggestion of proper activities to tryout for extra-curricular interests	13
4	Establishing friendly contact with faculty for future use	22
5	General discussion, encouragement and assistance with problem of self-confidence	155
6	Suggested treatment for specific personality difficulties	28
7	Discussion of meeting people and making friends	20
8	Treatment suggesting special things to do	11
9	Suggesting discussion between family and friends over mutual conflicts	25
10	Letter or interview with members of family over conflicts	28
11	Discussion of worries and other emotional problems	108
12	Advise transfer to another school because of home environment	12
13	Recommend welfare agency to assist student	2
14	Referral to Y M, C A for aid in social adjustment of student	19
15	Referral to psychiatrist for diagnosis.	53
16	Referral to Speech Clinic for diagnosis and treatment of speech difficulties.	31 541
Health and Physical Disabilities		
1	Discussion of handicaps	105
2	Advising remedial gym	8
3	Discussion of living arrangement sleep, diet, etc	4
4	Athletics suggested for better health	2
5	Referral to child health clinic	1
6	Referral to Health Service for special health examination	2 122
		6599

ANALYTICAL DESCRIPTION OF STUDENT COUNSELING

records, recommendation of a non-college type of vocational training, and recommendation that the student change his course of study were other procedures with high frequency.

The tabulation of the procedures used with social-personal-emotional problems indicate that the counselor relied on either a general discussion for encouragement and assistance with the problem of self-confidence or discussed worries and other emotional problems. The next most frequent step was to refer the student to the psychiatrist for diagnosis.

Discussion of the physical handicaps involved was by far the main method used with health problems. The counselor's discussion did not impinge upon medical advice but rather upon the relationship between physical condition and educational and vocational adjustments.

Summary

A general description based on 2053 cases was presented as a basis for analytical description of counseling in the Testing Bureau of the University of Minnesota over a four year period from 1932 to 1935 inclusive.

This description enables us to see how well the Bureau's counseling service is coordinated with the general personnel program of the University. By broad delineations of problem areas handled, of sources and amounts of data used and of procedures followed, the authors hope to break ground for a much needed basic description of the psychological processes involved in counseling interviewing conducted in a non-psychiatric guidance clinic for college students.

What is needed is like descriptions by other counseling services which may be based on the same or other philosophies of counseling. Such an accumulation of data should lead to even more specific descriptions in which recorded interviews would probably provide the raw data. Ultimately, it should be possible to determine experimentally which counseling procedures are most effective with what types of problems. This is the objective of evaluative research in the field of counseling.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

REFERENCES

- 1 Lindquist, E. F. *Statistical Analysis in Educational Research*. New York: Houghton Mifflin Company, 1940. 266 pp.
- 2 *Problems of College Education*. Edited by Earl Hudelson. Minneapolis: University of Minnesota Press, 1928. 449 pp.
- 3 Williamson, E. G. "Faculty Counseling at Minnesota," *Occupations*, XIV (1936), 426-433.
- 4 Williamson, E. G. and Bordin, E. S. "The Evaluation of Vocational and Educational Counseling: A Critique of the Methodology of Experiment," *Educational and Psychological Measurement*, I (1941), 5-24.
- 5 Williamson, E. G. and Bordin, E. S. "A Statistical Evaluation of Clinical Counseling," *Educational and Psychological Measurement*, I (1941), 117-132.
- 6 Williamson, E. G. and Bordin, E. S. "Evaluating Counseling By Means of a Control Group Experiment," *School and Society*, LII (1940), 434-40.
- 7 Williamson, E. G. and Dailey, J. G. *Student Personnel Work*. New York: McGraw-Hill Book Company, Inc., 1937. 313 pp.
- 8 Williamson, E. G., Longstaff, H. P., and Edmunds, J. M. "Counseling Arts College Students," *Journal of Applied Psychology*, XIX (1935), 111-124.
- 9 Williamson, E. G. and Sarbin, T. R. *Student Personnel Work in the University of Minnesota*. Minneapolis: Burgess Publishing Company, 1940. 115 pp.

A COMPOSITION TEST FOR FOREIGN LANGUAGES

LAWRENCE ANDRUS

University of Chicago

THIS PAPER discusses a type of French composition test, developed at the University of Chicago, which has in practice yielded remarkably good results as a measuring instrument and has proved very stable, i.e., has given comparable scores from year to year.

The test was developed as a part of the comprehensive examination in French 104-105-106, the sequence in Intermediate French given in the College of the University of Chicago. The College, as the term is used at the University of Chicago, includes the years corresponding to the freshman and sophomore years of the traditional four-year program. The prerequisites for admission to French 104-105-106 are two units of high school French or the successful completion of French 101-102-103, the sequence in Elementary French. Students in the College, as contrasted with more advanced students who desire to offer French 104-105-106 as an elective in a field related to their major field, may gain credit for the sequence only by passing the comprehensive examination given at the end of the Spring Quarter. The great majority of students in the course are College students. Since these students pass or fail solely on the basis of the comprehensive examination, the staff of the course and the examiner attempt, in every possible way, to make the examination as valid, as reliable, and as discriminating as they can. In the attempt to secure greater reliability, objective questions, or questions which can be scored with high objectivity, have been devised.

The Announcement of the College for 1940-41 describes French 104-105-106 thus: "The primary objective of the sec-

ond-year sequence is the standardization of the language abilities. To that end there is continuous training in formal and informal written and oral expression, aural comprehension and the accurate determination of the value of the printed word. Approximately twenty-five hundred pages are read, with reports, following individual programs.¹ This statement is a fair description of the course as given in the preceding four years, the period covered by this investigation. The type of test here discussed was intended to measure the outcomes of training in written expression. It was first used experimentally in the comprehensive examination of June, 1934.¹ In substantially its present form, it was included as a part of the 1935 examination, and retained in the following years with minor changes in the physical presentation.

The essential features of this type of test are as follows: a French passage is chosen which, in the judgment of the staff and the examiner, contains material suitable for testing at the level of the course, from the point of view of both vocabulary and syntax. It should be emphasized that the choice of an appropriate passage is extremely important, if the test is to yield maximal results. It may be necessary to read many pages before a suitable passage is located. This passage is then translated into good English. The next step is to go through the French text and delete certain words and phrases. The corresponding parts of the English translation are underlined and numbered to agree with the numbers replacing the omitted words and phrases in the French passage. The student is required to complete the French passage in accordance with the English translation. He is guided in this task by the numbers and the underlining.

A sample taken from the June, 1939, examination will give a better idea of the physical arrangement of the test than a lengthy explanation.

¹See Ernest Haden and John M. Stalnaker, "A New Type of Comprehensive Foreign Language Test," *The Modern Language Journal*, XIX, 2 (November, 1934), 81-92.

COMPOSITION TEST FOR FOREIGN LANGUAGES

Translation of French Text on Opposite Page

The old marquis de la Tour-Samuel, (1) *eighty-two years old*, arose and came (2) *to lean against the mantelpiece*. He said (3) *in his* (4) *somewhat trembling voice*

"(5) *I, too, know a strange thing, so strange* (6) *that it has been the obsession of my life* (7) *It is now fifty-six years* (8) *since this adventure* (9) *happened to me, and* (10) *a month doesn't go by* (11) *without my seeing it again in a dream* (12) *There has remained to me from that day a mark, an imprint of fear, do you understand me? Yes,* (13) *I underwent horrible fright,* (14) *for ten minutes,* (15) *in such a way that since that hour* (16) *a kind of constant terror* (17) *has remained* (18) *in my soul* (19) *Unexpected noises* (20) *make me start,* (21) *objects* (22) *that I make out* (23) *poorly in* (24) *the evening shadow give me* (25) *a mad desire* (26) *to run away*. Finally, I'm afraid (27) *at night*

"Oh! (28) *I shouldn't have admitted* (29) *that* (30) *before having arrived at my* (31) *present age*. Now I can say (32) *everything*. It is permitted (33) *not to be brave* before imaginary dangers, when (34) *you are eighty-two*. Before real dangers, (35) *I have never retreated, ladies*"

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Certain words and phrases have been omitted from the following French passage, and a number has been substituted for each omitted word or phrase. In each numbered space at the right, write in FRENCH the appropriate word or phrase. Be sure that your translation fits the French context. An English translation of the passage is given on the page opposite, the translation of each omitted word or phrase is underlined and preceded by a number which corresponds to the number in the French passage for each in reference. Note that there is not always exact correspondence in form between the French and the English. The old marquis uses the *conversational*, that is, *informal*, style, until he begins to tell his story, which is in *literary* style.

Le vieux marquis de la	(1)	
Four-Samuel (1), se leva	(2)	
et vint (2) la cheminée		
Il dit (3) sa voix (4)	(3)	(4)
— (5) sais une chose	(5)	
étrange, tellement étrange	(6)	(7)
(6) a été l'obsession de	(8)	(9)
ma vie (7) maintenant	(10)	(11)
cinquante-six ans (8)	(12)	(13)
cette aventure (9), et	(14)	(15)
(10) (11) en rêve (12)	(16)	(17)
de ce jour-là une marque,	(18)	(19)
une empreinte de peur,	(20)	(21)
me comprenez-vous? Oui,	(22)	(23)
(13) l'horrible épouvante	(24)	(25)
(14), (15) que depuis	(26)	(27)
cette heure (16) terreur	(28)	(29)
constante (17) (18), (19)	(30)	(31)
(20), (21) (22) je dis-	(32)	(33)
tingue (23) dans (24) me		
donnent (25) (26) J'ai		
peur (27), enfin		
Oh! (28) (29) (30) à		
mon âge (31) Maintenant		
je peux (32) dire Il est		
permis (33) devant les		
dangers imaginaires,		
(34) Devant les dangers		
véritables, (35), Mes-		
dames		

COMPOSITION TEST FOR FOREIGN LANGUAGES

A casual inspection of the sample suffices to reveal that this test form makes possible the use of a great variety of items, both as to content and as to length. The items may all be classified under one heading *usage*, with subclassification under *active vocabulary* (including idioms) and *grammar* (including syntax). By the proper choice of items tested, it is possible to vary at will both the level of difficulty and the proportion of vocabulary items and grammar items. In this way, validity with reference to specific objectives and content of a given course of study may be *built into* the test. For instance, in French 104-105-106 at the University of Chicago, a common practice has been to restrict items used in this test to the 2,500 words of highest frequency in the Vander Beke *French Word Book*². A similar procedure may be followed with respect to idioms, by using the Cheydleur *French Idiom List*³. Note that both upper and lower limits may be adopted. At present, grammar items must be validated on the basis of text books used and the subjective judgment of the instructing staff. When the *French Syntax Count*, begun under the direction of the late Professor Coleman, and now proceeding under the direction of Professor Keniston, is finally available, there will be an objective criterion of difficulty for French syntactical constructions. In Spanish, this invaluable aid has already been published⁴.

Theoretically, the most discriminating type of item for use in an achievement examination is one answered correctly by 50 per cent of the group taking the examination⁵. In practice, we almost never find a test containing even a majority of items of this type, except in the case of standardized tests which have been refined by statistical procedures, and even then,

²George E. Vander Beke, *French Word Book* (New York: Macmillan, 1929).

³Frederic D. Cheydleur, *French Idiom List, Based on a Running Count of 1,183,000 Words* (New York: Macmillan, 1929).

⁴Hayward Keniston, *Spanish Syntax List* (New York: Henry Holt & Co., 1937).

⁵Thelma Gwinn Thurstone, "The Difficulty of a Test and its Diagnostic Value," *The Journal of Educational Psychology*, XXIII, 5 (May, 1932), 335-343.

perhaps only with reference to the group on which the test was standardized. The classroom teacher interested in using the kind of test herein discussed would obviously have neither the time nor the statistical knowledge to go through the various steps required to develop a test composed largely of the most discriminating items. A fair approximation can, however, be attained by remembering that items that will be passed by practically all the group of students, or by almost none, have very little value for discrimination. They might be called "dead wood." The experienced teacher and his colleagues can, by subjective judgment, identify many such unprofitable items. Repeated use of a given test form and inspection of the results (not necessarily involving a formal item analysis, although that is always desirable when practicable), will tend to bring the teacher's subjective judgment of the worth of an item closer to an objective evaluation. It goes without saying that the value of an item as regards discrimination varies with the level of instruction and the content and method of the course, and should always be estimated in terms of these latter. To take a hypothetical example, in one school an item involving a particular use of the definite article in French might be highly discriminating, whereas in a school in a neighboring town, using a different course of study and a different method, the students might have received so much drill on this particular point that an item involving it would be passed by practically every student, and, hence, be of very little value for discrimination.

If the passage chosen, although otherwise desirable, is judged to lack an adequate number of instances of a particular construction considered important by the instructing staff, it is frequently possible to add items involving this construction by making slight changes in the French passage. It may not even be necessary to change the English translation at all, after such revision has taken place. The vocabulary items can be controlled in like manner.

The scoring of this test can be made very objective. At the time the test is constructed, as complete a key as possible

COMPOSITION TEST FOR FOREIGN LANGUAGES

is prepared, preferably by the entire staff of the course. This key facilitates the work of the scorer, who must, however, know the language thoroughly. The scoring cannot be entrusted to clerks. Whenever the scorer meets a correct answer not included in the key, he adds it to the key. Even with the necessity for consideration of such answers, the scoring is very rapid. In syntactical items, minor errors in spelling and mistakes in accents are disregarded, provided that the student uses correctly the construction on which the item hinges. The scoring thus becomes nearly as objective as that of a multiple-choice test.

TABLE 1

COMPOSITION TEST — FRENCH 104-105-106				
	1937	1938	1939	1940
No. of items	112	100	100	100
No. of points	112	100	100	100
No. of points in comprehensive examination	545	495	485	485
Mean	52.09	51.75	45.96	54.52
Standard deviation	17.80	15.95	14.86	13.98
Reliability*	.94	.94	.92	.91
Standard error of measurement ($\sigma_m = \sigma_t \sqrt{1-r}$)	4.36	3.91	4.20	4.19
Correlation with entire comprehensive examination	.92	.91	.88	.90
No. of cases	44	60	78	52

*Estimated by Kuder-Richardson formula No. 20,

$$r_{tt} = \frac{n}{n-1} \left(\frac{\sigma_t^2 - npq}{\sigma_t^2} \right)$$

Table 1 shows the main results of a statistical analysis of the different forms of this test used in the comprehensive examinations in French 104-105-106 at the University of Chicago during the four-year period 1937-1940 inclusive.

We note that the mean score in all four years was in the general neighborhood of 50 per cent of the possible number of points in the test. This is equivalent to saying that the *average* item was answered correctly by about 50 per cent of the group taking the examination. In 1940, a few items were

*See G. F. Kuder and M. W. Richardson, "The Theory of the Estimation of Test Reliability," *Psychometrika*, II, 3 (September, 1937), 151-60.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

purposely included which seemed *a priori* rather easy for the group tested, in order to test the effectiveness of such *a priori* judgment. These items were answered correctly by most of the students, and are reflected in the higher mean score for the 1940 test.

In each of the four forms of the test, the standard deviation was large enough so that the students' scores were well spread out, thereby facilitating classification of the students in rank order of merit. The differences in size of the standard deviation from one year to another are no greater than differences often found in administering the same test to two different groups of students, one slightly more homogeneous than the other. The spread of students' scores on this test is thus quite comparable from year to year.

For a test of 100 items, a reliability of .90 is commonly considered good. The lowest reliability coefficient estimated for the four-year period was .91 for the 1940 form, the highest, .94 for the 1938 form (this is relatively better than .94 for the 1937 form, since the latter contained twelve more items).

In none of the four years is the standard error of measurement as large as one-tenth of the mean score, and in none is it as large as one-third of the standard deviation. These values are satisfactorily low. They indicate that chance error in measurement has been kept within reasonable limits. Note that the difference between the highest and lowest standard errors of measurement here reported is only .45. To illustrate the meaning of this slight difference between the two extremes, let us assume that a student in 1937 and a student in 1938 each have a score of 40.00. The chances are two out of three that the true score of the 1937 student lies between 35.64 and 44.36, and that the true score of the 1938 student lies between 36.09 and 43.91.*

In only one year, 1939, does the correlation of students' scores on the composition test with their scores on the entire

**Editors' Note.* It will be recognized that not all statisticians would agree as to the validity of this interpretation.

COMPOSITION TEST FOR FOREIGN LANGUAGES

comprehensive examination fall below 90. The 1939 distribution includes one case (that of a student not registered for the course) which is so unlike the other cases in the group that it lowers the correlation by at least .01. It is customary to regard the correlation of a sub-test with the entire examination as spuriously high, since the sub-test is being correlated with a test of which it forms a part. In each of the four years, however, the composition test represents only about one-fifth of the total number of points in the entire comprehensive examination, and yet the correlation of this part with the entire comprehensive examination remains uniformly high, although a good part of the remaining material in the comprehensive examination has changed in character from year to year. This phenomenon, taken in connection with the relatively constant mean, standard deviation, reliability, and standard error of measurement, leads to the conclusion that under the conditions prevailing in French 104-105-106 at the University of Chicago this composition test represents a particularly stable type of measuring instrument.

Everyone must agree that the best method of testing French composition would be to require the student to write a free composition in French, if such tests could be scored reliably. Unfortunately, reliable scoring is in practice very hard to obtain. In the few instances where moderate success has been achieved, the process requires a great deal of time and involves essentially using the services of a jury of experts. Most teachers would probably agree that, at least at the lower and intermediate levels, the two elements which would assume the greatest importance in their judgment of a free composition in French are *active vocabulary* (including idioms) and *grammar* (including syntax). This type of test is capable of measuring students' achievement in these two elements reliably and objectively. The writer feels that *at the lower and intermediate levels* it is wiser to use a test which can do this than to run the risk of unreliable measurement which use of free composition entails.

The results of the statistical analysis shown in Table I

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

can be accepted unquestionably only for French 104-105-106 at the University of Chicago. There is no guarantee, but there is a strong presumption, that a test of this kind, constructed elsewhere with equal care, and with due attention to the objectives, content and method of the course of study, would yield equally favorable results. The technique could certainly be applied to Spanish, Italian, and Portuguese as well as to French, the sentence structure of German might prevent the technique from being as effective in that language as in the Romance languages.

PERFORMANCE TESTING IN PUBLIC PERSONNEL SELECTION

PART II¹

SIDNEY W. KORAN

Employment Board, Pennsylvania Department of Public Assistance

The Test for Graphotype-Addressograph Operators

THE POSITION of graphotype-addressograph operator occurs in each of the four regional financial offices of the Department of Public Assistance. Since these offices were conveniently located about the State and were equipped with a sufficient number of graphotype and addressograph machines to insure rapid and efficient conduct of the performance test, each of the four was used as an examination center and the examinees were permitted to appear at the one of their choice.

To minimize difficulties likely to arise because some examinees might be unfamiliar with the particular models on which the test was to be given, the notification form sent to each examinee included the statement "The examination to which you have been assigned has been designed to test your ability to operate the Class 6300 Graphotype and the Class 2700 Addressograph."

The test consisted of the following two parts and was set up and scored so that much the greater emphasis was placed on Part I:

- I Embossing names and addresses on Addressograph plates with the Class 6300 Graphotype
- II Printing cards from the embossed plates with the Class 2700 Addressograph

Standardization of the procedure was achieved by (1) having all tests administered under the supervision of trained

¹Part I of this article appeared in the July issue of this Journal

individuals, (2) requiring the examiners to repeat all directions to the examinee verbatim from the copy, (3) providing each examinee with a set of instructions setting forth the nature of the examination he was about to take and exactly what was expected of him, (4) using a stop watch for all timing, and (5) careful mechanical inspection of every machine at the beginning of each period of testing.

About 10 minutes before he was assigned to a machine each candidate was given a copy of the Instructions to Examinees (see Exhibit G) and told to read them carefully and to refer to them as often as he wished. When his turn arrived he was assigned to a Graphotype, furnished with a file drawer containing 22 plates and 20 frames, and given five minutes to familiarize himself with the machine and to practice embossing two of the plates. At the expiration of the practice period the examiner collected the two practice plates, furnished the examinee with the list of names and addresses to be embossed, and read aloud the appropriate sections of the Instructions to Examinees (see Exhibit H). At the expiration of 10 minutes the examinee was directed to stop embossing and to place the completed plates into frames. He was then assigned to an Addressograph and required to print a card from each embossed plate. If the examinee was unable to operate the Addressograph sufficiently well to print a legible copy from each plate, the examiner had an assistant print the plates on a strip of paper so that a record of the candidate's performance on the Graphotype would be available for scoring.

As with the Telephone Operator test (described in Part I of this article), the scoring procedure was designed to eliminate those whose performance fell below the standard established as the minimum acceptable, and to produce quantitative ratings reflecting the relative operating ability of those who survived that elimination. In establishing the qualifying point, consideration was given to (1) the requirements of the job, (2) the calibre of the individuals employed and available for employment, and (3) data on the agency's previous

PERFORMANCE TESTING IN PUBLIC PERSONNEL

experiences with a performance test for this type of position

The procedure followed in scoring the plates of those who met this criterion took into consideration both the speed and the accuracy with which the plates had been embossed. The examinee's raw score on the Graphotype portion of the test was determined by subtracting his error score—computed by counting each deviation from the key² as one error—from the total number of strokes completed. Examinees who demonstrated ability to operate the Addressograph were given additional credit up to 10 per cent of the total allowed for the complete performance test. Keys were constructed which reduced the scoring task to a routine operation.

The Test for Tabulating Machine Operators

The position of tabulating machine operator (IBM equipment) occurs only in the operating agency's State office. In administering the test in cities other than Harrisburg it was therefore necessary to arrange to use the facilities of the IBM Service Bureau.

To discourage individuals whose entire practical experience had been confined to the operation of sorters, numeric accounting machines, or Powers equipment from reporting for the performance test, the following statement, intended as a reminder, was included in the notification form sent to all examinees: "As previously indicated, the examination for this position has been designed to test your ability to operate both the IBM horizontal counting sorter and the IBM alphabetic accounting machine."

The test consisted of a two-part exercise. In Part I the examinee was required to (1) wire the plugboard of the alphabetic accounting machine for listing and for totals, (2) use the horizontal sorter, and (3) adjust and operate the alphabetic accounting machine so that certain alphabetic and numeric data from previously punched cards would be listed the same way as the data shown on the specimen form pro-

²Examples of deviations penalized were (1) incorrect letter or number, (2) spacing error of any kind, including line space, (3) insertion of a letter or number, and (4) omission of a letter or number.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

vided with the Instructions to Examinees (Exhibit I) Part II of the exercise was an extension of Part I in that it required the examinee to list specific data from the tabulating cards after (1) wiring the same plugboard to provide for numeric control and subtotals, (2) re-sorting the punched cards, and (3) making several additional adjustments on the accounting machine

About 10 minutes before the candidate was required to start the actual test, he was provided with a copy of the Instructions to Examinees and told to read them carefully and to refer to them as often as necessary throughout the examination. Attached to the sheet of Instructions were a sample punched card and a specimen sheet showing the form in which the data were to be listed by the alphabetic accounting machine in Part I of the exercise. A sample card and a portion of the specimen form sheet are reproduced as Exhibits J and K respectively

As soon as the examinee was ready to start the test he was provided with a plugboard, an adequate supply of the various sizes of wires needed in making the connections, and, for reference purposes, a type-bar layout and a plugboard diagram. The examinee was then reminded that the time limit for the entire test was one hour and forty-five minutes

Elapsed time was recorded by means of an electric job clock by stamping the starting time and finishing time of each operation on the examinee's job card. Since it was easier to secure the use of plugboards than tabulating and accounting machines, the equipment at each center included about three times as many plugboards as it did pieces of mechanical equipment. As a result of this it was sometimes necessary for an examinee to wait a few minutes for his turn at an alphabetic accounting machine. At no time, however, was it found that he was required to wait longer than 10 minutes, and this "lost time" was, of course, automatically taken care of by the job clock timing method employed. The small delays caused by having several times as many examinees wiring plugboards as could be accommodated at the machines were so

PERFORMANCE TESTING IN PUBLIC PERSONNEL

slight as to cause virtually no inconvenience to the candidates. On the other hand, the saving in time and expense which resulted from following this procedure was considerable.

When the examinee had finished wiring the plugboard to his satisfaction, one of the examiners inspected it to make certain that no connections had been made which would be likely to damage the machine. The examinee was then given a pack of 35 punched tabulating cards and directed to continue with Part I of the test. No limit was placed on the number of sheets of paper the examinee may have found it necessary to use nor on the number of times he was permitted to make changes in the plugboard wiring or machine adjustments. He was told to write his identification number on each sheet and to write "final copy" on the one he wished to submit for scoring. When Part I had been completed, the examinee continued immediately with Part II.

During the test the examiners made no attempt to rate the candidates on such points as the correctness of their particular approach, the acceptability of their work habits, nor, as already mentioned, the number of times they found it necessary to change the wiring or readjust the machine. The only factors taken into consideration in scoring the test were (1) the accuracy with which the assignments had been carried out, as shown by the finished products, and (2) the length of time consumed by the examinee in completing both parts of the exercise.

Reproduced as Exhibit L is a copy of the rating form which has been marked to show the scores of a typical examinee. The number in the parentheses after each item on the rating form is the maximum score obtainable for that item. An examinee completing both parts of the exercise correctly within 45 minutes would receive 30 points for Part I, 30 points for Part II, and 40 points for finishing within the minimum time bracket, making a total score of 100. A functional breakdown of the 60 points assigned to Parts I and II shows the following: sorting, 10 points, location of alpha-

betic and numeric fields, 30 points, totals and subtotals, 20 points.

The Test for Duplicating Machine Operators

The position of duplicating machine operator occurs in the State offices of the operating agency and the merit system agency. Persons filling these positions are required to operate several models of mimeograph and multilith machines. As most of the work involves the continuous use of the Class 1200 multilith and the Model 100 mimeograph in the performance of a large variety of duplicating jobs, and persons who can satisfactorily operate these models ordinarily have very little difficulty operating the older and less complicated types of equipment, the performance test was built around these particular machines, and the examinees were so informed well in advance of the date of the examination.

A copy of the Instructions to Examinees (Exhibit M) was furnished each candidate at least 10 minutes before he was required to begin the test. He was told to read these Instructions carefully and to keep them with him for reference throughout the examination.

The test for mimeograph operators was administered first. Each examinee was provided with a mimeograph stencil into which a solid box of typewritten material measuring $2\frac{7}{8}$ " by $4\frac{3}{4}$ " had been freshly cut, 75 $4\frac{1}{4}$ " by 6" white file cards, and 75 sheets of letter-size mimeograph paper on which a $3\frac{1}{4}$ " by $5\frac{1}{8}$ " frame had been printed. The examinee was then referred to his printed instructions which directed him to (1) place the stencil on the cylinder, (2) adjust the machine and duplicate 25 cards so that the material which had been typed on the stencil was centered on each card, (3) readjust the machine and duplicate 25 sheets so that the typed material was centered within the preprinted frame, and (4) remove the stencil and prepare it to be filed for future use. The frame preprinted on the letter-size paper was located in a position which required maximum adjustment of the machine's margin guides before the typewritten material on the stencil could be made to print within the required borders.

PERFORMANCE TESTING IN PUBLIC PERSONNEL

When he was ready for the multilith portion of the test, the candidate was provided with a photographic multex plate containing typewritten material measuring 6" by 8", 75 sheets of letter-size bond paper on which a 6½" by 8½" frame had been printed, and a supply of platex, keepeze, blankrola, repelex, and absorbent cotton. He was then referred to his printed instructions which directed him to (1) apply the platex, (2) put the plate on the machine, (3) adjust the machine and duplicate 25 sheets so that the typewritten material on the plate was centered within the preprinted frame, (4) remove the plate and prepare it to be filed for future use, and (5) clean the blanket. Two forms of the multilith plate were used alternately. These forms differed from each other only in the location of the typewritten material on the plate and were designed so that, while all candidates were required to make exactly the same kind of adjustments to center the material properly, the examinees were relieved of the necessity of setting up the machine after each run.

The examinees observed the candidates from a reasonable distance throughout the test in order to complete the rating form shown as Exhibit N. Timing was accomplished with a stop watch, and considerable use was made of the remarks column of the rating sheet to record all occurrences likely to contribute toward a fair evaluation of the examinee's performance. Whenever necessary, candidates were told how to turn on the particular model of equipment on which the test was given. Certain other bits of information, such as the use of the ink rolls on the multilith and the side margin adjustments on both the multilith and the mimeograph were also given, but careful note was made of the circumstances in each case so that the stipulated penalties could later be subtracted.

The half hour time limit on each machine was mentioned in the Instructions to Examinees but was not particularly emphasized. As the candidate started each part of the test the examiner said, "You will be allowed up to 30 minutes to complete this part of the test. The time you actually consume will enter into the computation of your score, but you

ought not to work so rapidly that the quality of your work suffers " Very few candidates required more than 15 minutes to complete the mimeograph test or more than 20 minutes for the multilith test, and most of those requiring more than this time were examinees who either got off to a bad start or were obviously so unfamiliar with the equipment that they were simply persevering while hoping for a miracle to occur. Persons in the latter group were encouraged to continue as long as they did not endanger the equipment. The additional time required to do this was cheerfully charged to public relations when it was discovered that, instead of rationalizing that if they had been given more time they would have succeeded, these candidates eventually insisted on withdrawing of their own accord and almost invariably thanked the examiners for "being so patient with me and giving me every break "

Scoring was accomplished by determining the number of points earned by the examinee for correctly accomplishing each of the items listed in the schedule of credits (Exhibit O), and then entering the appropriate amounts in the spaces provided on the summary sheet (Exhibit P). As finally worked out, the schedule of credits provided a weight of 60 for the multilith portion of the test, and 40 for the mimeograph.

In establishing the number of credits to be allowed for the successful accomplishment of each "item" of the test, consideration was given to the relative difficulty of the particular function under consideration. Thus, in scoring the mimeograph portion of the test, twice as much credit (8 points) was granted when the examinee's finished product presented evidence of correct side-margin adjustments as when the vertical margins were satisfactory (4 points). For the multilith, on the other hand, more credit (10 points) was given for proper adjustment of the side margins than for having the vertical margins correct (6 points).

PERFORMANCE TESTING IN PUBLIC PERSONNEL

Exhibit G

COMMONWEALTH OF PENNSYLVANIA
EMPLOYMENT BOARD

OF THE
DEPARTMENT OF PUBLIC ASSISTANCE

Harrisburg

PERFORMANCE TEST — SERIES 1900

GRAPHOTYPE AND ADDRESSOGRAPH MACHINE OPERATORS

July 1940

INSTRUCTIONS TO EXAMINEES

Important Failure to follow instructions may
result in disqualification from the examination

Study these instructions carefully. When you are ready to begin the examination, signal the Examiner. He will assign you to a machine and furnish you with the material with which you are to work.

Graphotype Machine

The examination for this machine consists of embossing a number of names and addresses in accordance with the form shown in the attached sample.

As soon as you have been assigned to a machine, the Examiner will furnish you with a file drawer containing 22 plates and 20 frames. You will be given 5 minutes to familiarize yourself with the machine during which time you may use 2, and only 2, of the plates to practice embossing.

At the conclusion of the practice period the Examiner will collect the 2 practice plates and give you a mimeographed list of names and addresses which you are to begin embossing as soon as he gives the signal to "Start."

Continue embossing until the Examiner calls "Time." *Do not put the plates in the frames as they are embossed, you will be required to do that later.*

"Time" will be called at the end of exactly 10 minutes.

The list of names to be embossed has purposely been made longer than even the fastest operators are likely to be able to complete. If the Examiner calls "Time" while you are in the midst of embossing a plate, you may remove the unfinished plate from the machine, but you *must not continue to emboss it.*

Inserting Plates in Frames

As soon as the Examiner tells you to do so, place each embossed plate in the lower part of a frame and arrange all frames in the file drawer so that they will be ready to run through the Addressograph.

When you have finished this task, the Examiner will provide you

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

with an envelope containing twenty 4" by 6" cards (on which your Identification Number has been printed) and assign you to an Addressograph

Addressograph Machine

The examination for this machine consists of printing from each plate that you have embossed

The machine will be set to print consecutively, and you will be required to make a single impression on each 4" by 6" card in the position shown on the attached sample.

PLACE YOUR APPOINTMENT SLIP, THE INSPECTION
SHEET, AND THE TWENTY CARDS INTO THE
LARGE ENVELOPE AND SEAL IT

PERFORMANCE TESTING IN PUBLIC PERSONNEL

Exhibit H
COMMONWEALTH OF PENNSYLVANIA
EMPLOYMENT BOARD
OF THE
DEPARTMENT OF PUBLIC ASSISTANCE
Harrisburg

PERFORMANCE TEST
GRAPHOTYPE AND ADDRESSOGRAPH MACHINE OPERATORS
SERIES 1900
July 1940

INSTRUCTIONS TO EXAMINER

- 1 Read the **INSTRUCTIONS TO EXAMINEES** and become entirely familiar with their contents before attempting to administer the examination
- 2 Examinees will be scheduled at the rate of three or four per hour where one set of machines is available, and at the rate of six or eight per hour where two sets are available
- 3 Do not admit anyone without an Admittance Slip unless he can establish his identity as an examinee who has qualified for the machine test
- 4 Provision should be made for the examinee to be comfortably seated away from the scene of the examination while he is awaiting his turn to operate the machines.
- 5 About 10 minutes before the examinee is assigned to a machine, take his fingerprint, hand him a copy of the mimeographed **INSTRUCTIONS**, and tell him to read them carefully. If he asks any questions, you may answer them, but it should not be necessary to furnish any information beyond that already appearing in the **INSTRUCTIONS**
- 6 When the examinee is ready to begin the examination and a Graphotype machine becomes available, assign him to the machine and furnish him with a file drawer containing 22 *plates* and 20 *frames* (all in perfect condition).
- 7 Tell the examinee he may have five minutes to familiarize himself with the machine and may practice embossing on two of the plates
- 8 At the expiration of five minutes (or before, if the examinee says he is ready to begin) collect the two practice plates and hand him the list of names to be embossed. Then say
"DO NOT BEGIN UNTIL I GIVE THE SIGNAL
EMBOSS EACH NAME AND ADDRESS ON A SEP-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

ARATE PLATE USING THE SAME FORM AS IN THE SAMPLE ATTACHED TO YOUR INSTRUCTIONS COPY THE NAMES AND ADDRESSES EXACTLY AS THEY APPEAR, *OMITTING ALL PUNCTUATION* AT THE END OF *10 MINUTES*, WHEN I CALL 'TIME,' YOU MUST STOP EMBOSSING "

- 9 Say to the examinee. "READY, *START* "
 - 10 Permit the examinee to continue exactly *10 minutes*. Then say, "TIME STOP EMBOSSING "
 - 11 If a plate is in the machine, permit him to remove it. Take away the list of names and all blank plates. Then say "PLACE EACH EMBOSSED AND PARTIALLY EMBOSSED PLATE INTO THE LOWER PART OF A FRAME AND ARRANGE THE FRAMES IN THE FILE DRAWER SO THAT THEY CAN BE RUN THROUGH THE ADDRESSOGRAPH "
 - 12 When the examinee has placed each plate in a frame, take away the unused frames, hand him his envelope containing the cards, and assign him to an Addressograph. Then say "PRINT THE CARDS FROM THE PLATES. MAKE ONLY ONE IMPRESSION ON A CARD AND IN APPROXIMATELY THE SAME POSITION AS SHOWN ON THE SAMPLE ATTACHED TO YOUR INSTRUCTIONS. THE ADDRESSOGRAPH HAS BEEN SET TO PRINT CONSECUTIVELY GO AHEAD "
 - 13 When the examinee has made an impression from each plate, tell him to place the *20 cards* (printed and unprinted) into the envelope together with his *Admittance Slip* and *Instructions* and seal the envelope.
- Note* If the examinee is unable to operate the Addressograph sufficiently well to print a legible copy from each plate he has embossed, have the plates printed on a strip of paper so that a record of his performance on the Graphotype will be available for scoring.
- 14 If at any stage of the machine operation you and the Addressograph representative are convinced that the examinee does not possess sufficient knowledge of the operation of either machine to continue with safety to himself and without damage to the equipment, the test may be halted. If this becomes necessary, a full statement of the circumstances must be written on the back of the Admittance Slip and signed by both you and the representative. In any instance in which the plates themselves will be valuable as possible exhibits, they should be enclosed in the examinee's envelope before sealing.

PERFORMANCE TESTING IN PUBLIC PERSONNEL

Exhibit I

COMMONWEALTH OF PENNSYLVANIA
EMPLOYMENT BOARD
OF THE
DEPARTMENT OF PUBLIC ASSISTANCE

Harrisburg

PERFORMANCE TEST — SERIES 2100

SENIOR TABULATING MACHINE OPERATOR

December 1940

INSTRUCTIONS TO EXAMINEES

Important Failure to follow instructions may
result in disqualification from the examination

Study these instructions carefully. As soon as a machine becomes available, the Examiner will give you further instructions and furnish you with all necessary material.

The performance test for this position will consist of a two part exercise designed to determine your ability to operate the IBM Horizontal Counting Sorter and IBM Alphabetic Accounting Machine. *The time limit for the entire test is 1 hour and 45 minutes.*

You will be given 35 tabulating cards (into which various data have been punched) and a plugboard for the Alphabetic Accounting Machine. Note that the model of the Accounting Machine being used has 32 counters and 55 type bars of which numbers 19 to 43 are alphabetic.

The following operations should be carried out in the order indicated.

PART I

1. Wire the plugboard so that the machine will list the following information *exactly as shown on the attached sample*

Card Number

Name (last, first *initial*, middle *initial*)

Social Security Number

Total Benefits Paid

Weekly Benefit Amount

Reason

Note In addition to listing the data, the machine is to be wired to show totals at the end of each of the following fields

Total Benefits Paid (allow for six digits)

Weekly Benefit Amount (allow for six digits)

2. Sort the cards in "Card Number" order.
3. Write your Identification Number in the space provided on Form I and list the data from the cards to conform to the sample and the above instructions

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

PART II

- 1 Wire the plugboard to control on "Reason "
- 2 Sort the cards by "Reason," disregarding "Card Number "
- 3 Set automatic hammerlock control to eliminate the listing of names.
- 4 On another copy of Form I write your Identification Number in the space provided and list the data from the cards, single spaced, to show subtotals for each reason (for Total Benefits Paid and Weekly Benefit Amount).

PLACE YOUR ADMITTANCE SLIP, THIS INSTRUCTION SHEET, THE PUNCHED CARDS, AND BOTH COPIES OF FORM I IN THE MANILA ENVELOPE AND SEAL IT

PERFORMANCE TESTING IN PUBLIC PERSONNEL

Exhibit J

[illegible]

Exhibit B

PERFORMANCE TEST FOR SENIOR TABULATING MACHINE OPERATORS—DECEMBER 1940

Card Number	Name	Social Security Number	Total Benefits Paid	Weekly Benefit Amount	Reason
120	CARSONBERGER J J	567 89 3428	45365	7895	4

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Exhibit L

PERFORMANCE TEST—TABULATING MACHINE OPERATORS
Series 2100

<i>B 24356</i>	<i>Allegheny</i>	
Ident Number	Legal County	File Number
<i>Withdrawn from Examination (0)</i>		<u>✓</u>
PART I		
<i>Sorting by Card Number (5)</i>		<u>5</u>
<i>Location of Fields (15)</i>		
Card Number (1)		<u>1</u>
Name (2)		<u>2</u>
Social Security Number (2)		<u>2</u>
Total Benefits Paid (2)		<u>0</u>
Weekly Benefit Amount (2)		<u>2</u>
Reason (2)		<u>2</u>
Totals		
Benefits Paid Column (2)		<u>2</u>
Benefit Amount Column (2)		<u>0</u>
<i>Accuracy of Totals (10)</i>		
Benefits Paid Column (5)		<u>5</u>
Benefit Amount Column (5)		<u>5</u>
PART II		
<i>Sorting by Reason (5)</i>		<u>5</u>
<i>Location of Fields (15)</i>		
Card Number (1)		<u>1</u>
Name <i>eliminated</i> (2)		<u>2</u>
Social Security Number (2)		<u>0</u>
Total Benefits Paid (2)		<u>0</u>
Weekly Benefit Amount (2)		<u>2</u>
Reason (2)		<u>2</u>
Subtotals		
Benefits Paid Column (2)		<u>2</u>
Benefit Amount Column (2)		<u>2</u>
<i>Accuracy of Subtotals (10)</i>		
Benefits Paid Column (5)		<u>5</u>
Benefit Amount Column (5)		<u>5</u>
<i>Time (40)</i>		
45 minutes or less (40)	} <u>48 min</u>	<u>30</u>
46 to 60 minutes (30)		
61 to 75 minutes (20)		
76 to 90 minutes (10)		
91 to 105 minutes (0)		
TOTAL RAW SCORE		<u>82</u>
Scored by <u>S W K</u>		Checked by <u>P, N E</u>
(Form EB-742)		

PERFORMANCE TESTING IN PUBLIC PERSONNEL

Exhibit M

COMMONWEALTH OF PENNSYLVANIA
EMPLOYMENT BOARD
OF THE
DEPARTMENT OF PUBLIC ASSISTANCE
Harrisburg
PERFORMANCE TEST
DUPLICATING MACHINE OPERATORS
SERIES 2800
November 1940

INSTRUCTIONS TO EXAMINEES

Important Failure to follow instructions may
result in disqualification from the examination

Study these instructions carefully. As soon as a machine becomes available, the Examiner will give you further instructions and furnish you with all necessary material.

Mimeograph Machine Time limit, 30 minutes. The Examiner will furnish you with the following material:

- 1 newly cut mimeograph stencil
- 75 4" by 6" cards
- 75 sheets of pre-printed 8½" by 11" mimeograph paper
(sample attached)

The examination for this machine will consist of (1) putting the stencil on the cylinder, (2) adjusting the machine and duplicating 25 cards so that the material which has been typed on the stencil is centered on each card, (3) readjusting the machine and duplicating 25 sheets so that the typed material is centered within the pre-printed box, and (4) removing the stencil and preparing it to be filed for future use. (You may use as many cards and sheets of paper as necessary in setting up the machine, but do not waste any. All material will be considered in determining your score in the examination.)

WHEN YOU HAVE COMPLETED THIS PORTION
OF THE TEST, PLACE THE STENCIL AND ALL
USED CARDS AND PAPER INTO THE LARGE
MANILA ENVELOPE

Multilith Machine Time limit, 30 minutes. The Examiner will furnish you with the following material:

- 1 photographic multex plate
- 75 sheets of pre-printed 8½" by 11" bond paper (sample attached)
- Supply of Platex, Keepeze, Blankrola, Repelex, and absorbent cotton

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

The examination for this machine will consist of (1) applying Platem, (2) putting the plate on the machine, (3) adjusting the machine and duplicating 25 sheets so that the typed material is centered within the pre-printed box, (4) removing the plate and preparing it to be filed for future use, and (5) cleaning the blanket. (You may use as many cards and sheets of paper as necessary in setting up the machine, but do not waste any. All material will be considered in determining your score in the examination.)

WHEN YOU HAVE COMPLETED THIS PORTION OF THE TEST, PLACE YOUR ADMITTANCE SLIP, THIS INSTRUCTION SHEET, AND ALL *USED* PAPER INTO THE LARGE MANILA ENVELOPE AND SEAL IT.

PERFORMANCE TESTING IN PUBLIC PERSONNEL.

Exhibit N

COMMONWEALTH OF PENNSYLVANIA
EMPLOYMENT BOARD
OF THE
DEPARTMENT OF PUBLIC ASSISTANCE
Harrisburg

PERFORMANCE TEST FOR DUPLICATING MACHINE OPERATORS
SERIES 2800

November 1940

EXAMINER'S RATING SHEET

Examinee's Id No

Mimeograph Machine

OPERATION	Card	Paper	Remarks
Place stencil on cylinder		X	
Adjust paper feed			
Make side margin adjustment			
Make vertical margin adjustment			
Use of print recorder			
Run copies			
Take off stencil	X		
Clean stencil	X		
TIME	START	STOP	ELAPSED

Multilith Machine

OPERATION		Paper	Remarks	
Platex plate				
Put on plate				
Ink up				
Pull proof				
Locate form in proper position				
Clean blanket				
Set counter				
Run copies				
Clean plate				
Take off plate				
Keepeze				
Clean blanket				
TIME	START	STOP	ELAPSED	

Date_____

Examiner_____

Examiner_____

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Exhibit O

PROCEDURE FOR SCORING DUPLICATING MACHINE
OPERATOR PERFORMANCE TEST

SERIES 2800

SCHEDULE OF CREDITS

Note: The examinee's rating for each item is to be determined in accordance with the following schedule and entered in the appropriate space on the scoring sheet (Form EB-760). All ratings and totals must be checked by a second scorer.

Mimeograph (40)

Time (16)

- 1 to 10 minutes — 16 points
- 11 to 15 minutes — 12 points
- 16 to 20 minutes — 8 points
- 21 to 25 minutes — 4 points
- 26 to 30 minutes — no credit

Stencil (2)

- Removed and cleaned properly (credit if checked on rating sheet)
— 2 points

4" by 6" cards (11)

- Practice cards (not more than 15) — 2 points
- Number of copies (25 to 30) — 2 points
- Use of counter — 1 point
- Vertical margin adjustment (80% of final copies with at least $\frac{3}{8}$ " margin top and bottom) — 2 points
- Side margins (80% of final copies with at least $\frac{1}{8}$ " margin each side) — 4 points

8½" by 11" paper (11)

- Practice sheets (not more than 15) — 2 points
- Number of sheets (25 to 30) — 2 points
- Use of counter — 1 point
- Vertical margins (80% of final copies not touching horizontal lines) — 2 points
- Side margins (80% of final copies not touching vertical lines) — 4 points

Penalty (-4)

- No 1—If examinee was given information on side margin adjustment—*subtract 4 points* (but only when examinee has received credit for correct side margin adjustment)

Multilith (60)

PERFORMANCE TESTING IN PUBLIC PERSONNEL

Time (24)

- 1 to 10 minutes — 24 points
- 11 to 15 minutes — 18 points
- 16 to 20 minutes — 12 points
- 21 to 25 minutes — 6 points
- 26 to 30 minutes — no credit

Clean plate (credit if checked on rating sheet) — 5 points

Clean blanket (credit only if second listing is checked on rating sheet)
— 2 points

8½" by 11" paper (29)

Practice sheets (not more than 15) — 5 points

Number of copies (25 to 28) — 3 points

Use of counter — 2 points

Side margins (at least 3/16" on each side) — 6 points

Vertical margins (at least 3/16" top and bottom) — 10 points

Inking (3)

Evenness — 2 points

Blackness — 1 point

Penalties (-10)

No 2—If examinee was given information on use of ink rolls —
subtract 5 points

No 3—If examinee was given information on side margin adjustment — *subtract 3 points* (but only when examinee has received credit for correct side margin adjustment)

No 4—If examinee left an ink roll in contact with plate cylinder —
subtract 2 points

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Exhibit P

PERFORMANCE TEST—DUPLICATING MACHINE OPERATORS—SERIES 2800

File Number	County	Ident. Number
-------------	--------	---------------

Operation of Mimeograph (40)

Withdrew or was stopped by Examiners (F)

Time (16)

Stencil (2)

4" by 6" Cards (11)

Practice cards (2)

Number of copies (3)

Vertical margin adjustment (2)

Side margins (4)

8½" by 11" Paper (11)

Practice sheets (2)

Number of sheets (3)

Vertical margin adjustment (2)

Side margins (4)

Operation of Multilith (60)

Withdrew or was stopped by Examiners (F)

Time (24)

Clean Plate (5)

Clean Blanket (2)

8½" by 11" Paper (29)

Practice sheets (5)

Number of copies (5)

Vertical margin adjustment (10)

Side margins (6)

Inking (3)

Total

Penalties: No. 1___ No. 2___ No. 3___ No. 4___ Subtract

Total Raw Score

Scored by_____ Checked by_____

(Form EB-760)

THE VALUE OF INTELLIGENCE QUOTIENTS OBTAINED IN SECONDARY SCHOOL FOR PREDICTING COLLEGE SCHOLARSHIP

L. D. HARTSON

and

A. J. SPROW

Oberlin College

IN SELECTIVE admission to college, and particularly in the award of scholarships, it is the practice to request a report of scores made by the candidates in intelligence tests. This study reports (1) the relative value of these different tests for predicting (a) total high school scholarship,¹ (b) college freshman scholarship, and (c) seven-semester college scholarship; (2) the comparative validity of these tests and the Ohio State University Psychological Examination; (3) the average I.Q. of the student body, as determined by these various instruments; (4) comparison of the I.Q.'s of the Oberlin group with those of the Terman-Merrill standardization group; (5) average freshman scholarship for students of the different I.Q. levels.

A total of 835 freshmen entered the College of Arts and Sciences, Oberlin College, during the period, 1934 to 1940, for whom I.Q.'s were available, which could be identified with specific forms of tests, in groups large enough to warrant statistical treatment. In six cases there were two scores, making the total of 841 in Table 1. Of these, 253 had progressed as far as their eighth semester. For these, the computations are based upon the scholarship record for seven semesters (those

¹The figures used in the computations of high school scholarship represent, not the actual grades, but "credit points" obtained by a system used to equate different grading schemes.

TABLE 1

CORRELATIONS BETWEEN I. Q.'S DERIVED FROM VARIOUS TESTS AND (1) HIGH SCHOOL SCHOLARSHIP, (2) COLLEGE FRESHMAN SCHOLARSHIP, (3) OSU TEST SCORES, WITH (+) MEANS AND SIGMAS

Test	N	Scholarship		Fresh. Wk		Test Scores		H. S.		Scholarship ¹		Freshman	
		H. S.	Freshman	OSU Test	Means	Sigmas	M	σ	M	σ	M	σ	
Outs	444	.522	364	.504	120 00	8.85	75 24	18.06	49 06	28 10			
Terman	221	.281	403	.610	121 23	10.04	75 11	16 82	49.48	28 05			
H-N	110	.396	480	.544	121 86	10.39	77 15	18.31	49 05	27 71			
National	38	.212	.287	.487	127 21	12.06	82 18	12.47	54 18	24.41			
K-A	28	.247	.178	.456	124 96	12.20	72.21	20 11	53.71	30 48			
OSU (pre-entrance)	258	.365	474	.743	55.73	26.09	65.93	22 81	39.96	30 08			

¹High school scholarship is expressed in terms of "credit points" (see footnote, p. 1); college scholarship, in terms of proportional rank.

VALUE OF I.Q.'S FOR PREDICTING COLLEGE SCHOLARSHIP

used for the award of Phi Beta Kappa). In order to make the scholastic records equivalent for the several classes of varying size, scholarship is handled in terms of proportional class rank. Because the reports did not, in all cases, specify the particular variety of Otis test employed, all the Otis scores have been grouped. All the I.Q.'s here considered were derived from group tests.

Results

Table 1 reports the coefficients of correlations (1) between the I.Q.'s on the Otis, Terman, Henmon-Nelson, National, and Kuhlmann-Anderson Tests, and scores on the Ohio State University Psychological Examination administered before matriculation in college as one variable, and high school scholarship; (2) between the above-named tests and first semester college scholarship; (3) between I.Q.'s on one or another of the first five tests and scores on the OSU test, administered during freshman week, with the means and sigmas. In the case of the National and of the Kuhlmann-Anderson tests the N is rather small, and all the data, therefore, are less reliable than those obtained with the other tests. To obtain a basis for comparing the validities of the OSU test and each of the others, Table 2 reports, for each of the test

TABLE 2

CORRELATIONS BETWEEN THE OSU TEST SCORES AND SCHOLARSHIP OF THE GROUPS TESTED WITH THE OTIS, TERMAN, HENMON-NELSON, NATIONAL, KUHLMANN-ANDERSON AND OSU TEST (PRE-ENTRANCE GROUP) WITH MEANS AND SIGMAS

Test Group	N	Scholarship		Means	Sigmas
		High Sch	Freshman		
Otis	444	.394	.579	49.35	28.95
Terman	221	.337	.550	51.47	28.00
Henmon-Nelson.	110	.458	.604	48.59	29.22
National	38	.473	.631	54.18	25.36
Kuhlmann-Anderson	28	.633	.564	51.93	26.63
OSU Test (pre-entrance)	258	.510	.629	48.72	28.72

groups, the correlation between scores in the OSU test and (a) high school scholarship and (b) first semester college scholarship, with means and sigmas.

1. *Relative Validities of the I.Q. Tests.* A comparison of the validities of the different tests yielding I.Q.'s indicates that the Henmon-Nelson test gets first place. However, the relative deviate of the difference (k) between the correlation of Henmon-Nelson I.Q.'s and college scholarship (.480) and the corresponding coefficient for the Otis test (.364) is but 1.20.

2. *The Prediction of High School and College Scholarship.* The coefficients indicating the relationship between I.Q. and high school scholarship range between .212 and .396. With the exception of the Kuhlmann-Anderson test (for which there are but 28 cases) the I.Q.'s constitute a better basis for predicting college scholarship than they do high school grades. When validated against college scholarship, the coefficients range between .287 (omitting Kuhlmann-Anderson) and .480. Byrns and Henmon, who used I.Q.'s obtained from National Intelligence Tests, administered in the 4th to 8th grades, also found a closer correlation between I.Q. and first semester college scholarship (.454) than between I.Q. and total high school scholarship (.426) (2). Higher validities for the college criterion were also obtained for both of the OSU examinations. With the pre-entrance OSU Test the coefficients are .365 and .474, and with the Freshman Week test, they are .510 and .629, for high school grades and college scholarship, respectively. These results substantiate previous findings at Oberlin. For the 511 men and 609 women who entered as freshmen during the period, 1931 to 1934, the correlation between college scholarship and OSU Test intelligence is represented by coefficients of .605 and .574, for the men and the women, respectively, whereas the correlation between test intelligence and high school scholarship is represented by coefficients of .398 and .380 (3). It will be noted that the OSU Test scores have higher validity than does the I.Q., as indicated by the coefficients obtained

VALUE OF I.Q.'S FOR PREDICTING COLLEGE SCHOLARSHIP

TABLE 3

COMPARATIVE VALIDITIES OF THE OSU TEST AND THE I. Q. TESTS FOR
PREDICTING HIGH SCHOOL AND COLLEGE SCHOLARSHIP

Test Group	High School Scholarship		College Scholarship	
	OSU Test	I.Q.	OSU Test	I.Q.
Otis394	.322	.579	.364
Terman337	.281	.550	.403
Henmon-Nelson..458	.396	.604	.480
National473	.212	.631	.287
Kuh!mann-Anderson	.633	.247	.564	.178

when correlations were computed for each of the I.Q. populations between the scores made with the OSU Test and the two scholarship criteria (Table 3)

The OSU Test is designedly a more difficult one than the other tests. Although some tests have as many items as the OSU Test, none requires as much time. All of the I.Q. tests are time limited, maximum time being 30 minutes, whereas the OSU Test was administered by work-limit method, students usually taking at least two hours.

3 *Comparative Validity of Pre-entrance and Freshman Week OSU Test Scores.* That the higher coefficients obtained for the OSU Test may not be due entirely to its greater difficulty, however, is suggested by a comparison of the coefficients obtained for the OSU Test under two sets of conditions. There were 258 students who took the OSU Test some time before entering college who were re-examined with this test during their Freshman Week. (In some instances the same form of the test was used, but usually it was another form) The Freshman Week test yielded substantially higher validity figures with both criteria. .510 as compared with .365 for high school scholarship, and .629 as compared with .474 for college freshman scholarship These coefficients, with the means and sigmas, are reported in Tables 1 and 2. It will be noted that the group given the OSU (pre-entrance) Test had distinctly lower scholastic records than the others and that they also displayed greater variability. This is to be

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

explained by the fact that the great bulk of these students were given the test because their high school record made their admission questionable. On the other hand, the group contained some who, being of exceptionally high caliber, were applying for scholarships. In most cases, it may be presumed, the OSU Test was administered under conditions of greater motivation than prevailed with the I.Q. tests.

4. *Intercorrelations Between the Test Scores.* The intercorrelations between the scores made on the OSU Test and the different forms of I.Q. range from .456 to .610; the sequential order from the higher to lower coefficients being: Terman, Henmon-Nelson, Otis, National, and Kuhlmann-Anderson.

5. *Validation of Tests Against Total College Scholarship.* There are 253 students for whom scholastic grades are available for seven semesters of the college course. Because the numbers were too small to warrant separate computations for each of the tests, all of the I.Q.'s were combined and the validity coefficients computed, using both the one- and the seven-semester criterion. Validity coefficients were obtained for the OSU Test for the same population. These are given in Table 4. The two validity figures for the I.Q.'s are .341 and .319, and for the OSU Test scores the figures are .501 and .438. Although scores on the OSU Test are more valid bases for prognosing college grades than is the I.Q. at both levels, their superiority for predicting total college scholarship is less than when used for predicting freshman grades. From Table 4, one may also note that, as in the computations re-

TABLE 4

CORRELATIONS BETWEEN I. Q.'S AND OSU TEST SCORES AND (1) HIGH SCHOOL SCHOLARSHIP, AND SCHOLARSHIP FOR (2) ONE AND (3) SEVEN SEMESTERS; WITH MEANS AND SIGMAS

Test Score	N	High Sch.	1 Sem.	7 Sems.	Mean	Sigma
I. Q.	253	.171	.341	.319	121.24	10.00
OSU Test	253	.307	.501	.438	50.95	29.85
Mean		75.82	53.37	47.16		
Sigma		17.19	25.68	29.00		

ported for the other populations, the I.Q. (and OSU Test score) shows a closer relationship for college freshman scholarship than for high school scholarship

6. *Average I.Q. of the Oberlin Student Body.* The mean I.Q. of the 835 freshmen, as measured by these group tests, is 121.06. There is substantial agreement on this point between the Otis, Terman, and Henmon-Nelson tests (see Table 1). The cases measured by the National and Kuhlmann-Anderson tests, for which the means are 127.21 and 124.96 respectively, are too few to influence the general average materially. The average for the group of 253 who attained senior status is 121.24, thus indicating that practically no selection occurred between the freshman and the senior year in terms of I.Q. This is corroborated by the OSU Test standing of the freshman and senior groups. In terms of local freshman norms, the mean score of the 835 students is 49.97, thus indicating that they are an almost completely perfect sample of the Oberlin first-year population. The mean score of the 253 who became seniors is 50.95. The seniors do, however, constitute a somewhat selected group in terms of college scholarship. This is indicated by the fact that, whereas the mean freshman scholarship of the entire group is represented by a proportional rank of 49.55, the mean freshman rating of those who persisted until they reached senior status is 53.37—the larger figure represents higher scholarship status—and the mean scholarship of the 588 who had not become seniors is 47.91. The critical ratio of the difference between the freshman scholarship of those who became seniors and those who did not is 2.73.

7. *Comparison of Oberlin Students with the Terman-Merrill Standardization Group.* Figure 1 presents a graphic comparison of the Oberlin students with the normal group of 2904 used in the standardization of the Terman-Merrill Binet test (4, p. 37). The numbers and proportions of the Oberlin population at the different I.Q. levels are reported in Table 5. The I.Q.'s of the Oberlin group range from 92 to 169, 99 per

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 5
FRESHMAN AND SENIOR SCHOLARSHIP OF STUDENTS OF
DIFFERENT I. Q. LEVELS

I Q.	Entrants				Seniors				
	N	%	Scholarship Mean	Scholarship Range	N possible	N actual	%	Scholarship Mean	Scholarship Range
166-170	3	0.4	97.7	97-99	3	2	66.7	96.0	93-99
161-165	0				0				
156-160	0				0				
151-155	1	0.1	94.0	94	1	1	100.0	87.0	87
146-150	13	1.6	72.9	13-95	5	3	60.0	63.3	30-90
141-145	14	1.7	72.1	26-97	7	5	71.4	66.0	12-84
136-140	44	5.3	70.9	5-98	16	13	81.3	51.2	1-99
131-135	57	6.8	59.8	4-99	18	14	77.8	49.6	13-99
126-130	110	13.2	58.4	1-99	43	36	83.7	57.4	1-99
121-125	178	21.3	52.1	1-99	74	57	77.0	53.3	5-100
116-120	179	21.4	45.5	1-98	71	52	73.2	42.4	2-91
111-115	119	14.2	41.3	1-97	48	32	66.7	38.5	2-95
106-110	75	9.0	33.3	1-94	39	26	66.7	31.1	5-65
101-105	29	3.5	36.7	2-98	18	8	44.4	37.3	4-69
96-100	8	0.9	25.1	3-42	2	2	100.0	27.0	26-28
91-95	5	0.6	25.8	7-69	2	2	100.0	24.5	14-35
	835				347	253	72.9		

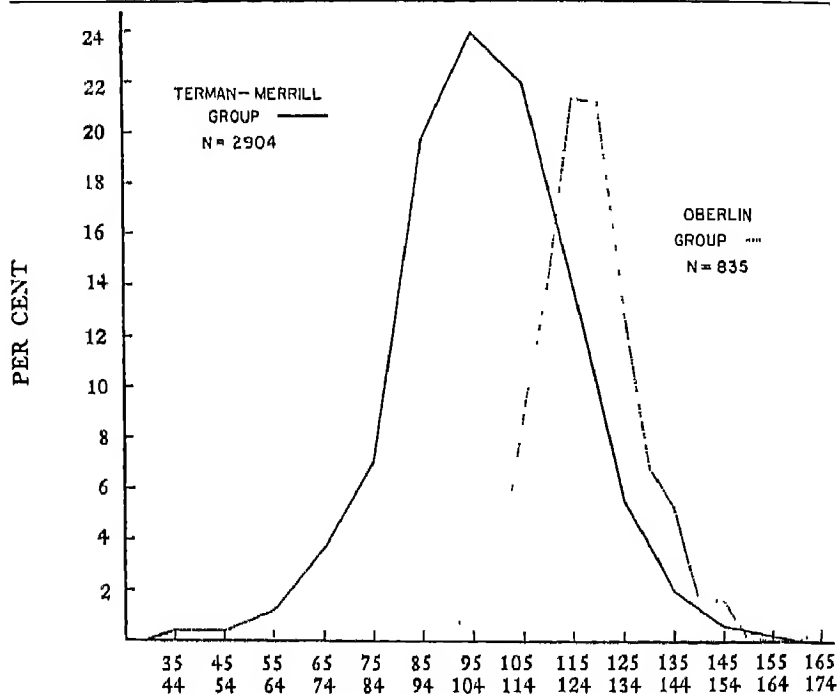


Figure 1
Distributions of the IQ's in the Terman-Merrill Standardization Group
and the Oberlin Group

VALUE OF I.Q.'S FOR PREDICTING COLLEGE SCHOLARSHIP

cent exceeding the Terman-Merrill mean. The Oberlin sample is rather sharply peaked, showing little kurtosis, but it displays a slight positive skewness. Variability is much less than that of the Terman-Merrill sample, sigma being 10.4 as compared with 16.4 for the larger group.

8. *Mean Scholarship of Students of Different I.Q. Levels.* Table 5 reports the numbers and proportions of students of the different levels of I.Q. with their (1) mean freshman scholarship rank, (2) mean senior scholarship rank, (3) the range of scholarship achievement for those at each level, and (4) the proportion of those in college long enough to have attained senior status who did so, for each I.Q. level. Examination of the table reveals the following salient facts:

(a) As indicated by the correlation coefficients previously noted, the general tendency is for those of higher I.Q. to make the better scholastic records.

However, (b) the range of scholastic performance is, with few exceptions, remarkably similar at each test level. Freshman achievement in the highest and the lowest deciles is recorded for students with I.Q.'s ranging all the way from 105 to 140, although no student with an I.Q. below 111 achieved a top tenth ranking for the entire college course. There was one student with an I.Q. of 105 who achieved a proportional rank of 98 in freshman scholarship and has been in the upper tenth of her class in each of the two subsequent years. Her centile score, according to state norms, on the OSU Test is, however, 71, so the later test is evidently a more accurate index of her intellectual ability.

(c) The four students with I.Q.'s above 150 all made exceptionally good records.

(d) Sufficient time has elapsed to permit but four students whose I.Q.'s are below 101 to become seniors. They have all obtained the A.B. degree, but in only one instance was this achieved in the normal four-year period. By persistent effort, however, they did finish the course, and all of them ranked above the lowest decile of their class. This is comparable with Adams' finding at the University of Texas.

(e) The retentive power of the college was not materially greater for those at the higher than for those at the lower extremes of the distribution

(f) Some degree of selectivity is indicated, however, by a comparison of those with I.Q.'s above 126 with those whose I.Q.'s are below 116. Of the 93 in the group with the higher I.Q.'s, 74, or 79.57 per cent, persisted, whereas but 70, or 64.22 per cent, of the 109 with the lower I.Q.'s persisted to the senior year. As the ratio of the difference in these proportions to the standard error of the difference is 2.47, it is fairly significant.

Summary

1. I.Q.'s were available for 835 entering freshmen and for 253 of these who had reached the senior year, the scores having been derived from the following group tests: Otis, Terman, Henmon-Nelson, National and Kuhlmann-Anderson. Scores on the OSU Psychological Examination were also obtained.

2. The difference in the power of the different I.Q. tests to predict college scholarship was not statistically significant.

3. The I.Q.'s constitute a better basis for predicting college grades than they do for prognosing total high school scholarship. This is also true of the OSU Test scores.

4. The OSU Test is more successful than any one of the other tests in predicting scholarship in high school as well as in college.

5. The OSU Test taken during Freshman Week correlates more closely with both secondary and college scholarship than does the same test taken during the senior year in high school.

6. The OSU Test predicts freshman scholarship better than it does total college scholarship.

7. The average I.Q. of the freshmen is 121. The average obtained by the Otis, Terman, and Henmon-Nelson tests is virtually the same. The averages for the small number tested with the National and Kuhlmann-Anderson tests are 127 and 125, respectively.

VALUE OF I.Q.'S FOR PREDICTING COLLEGE SCHOLARSHIP

8. The average I.Q. for the seniors is also 121. As the mean OSU Test score for the seniors is but one percentile point higher than that for the freshmen, it is evident that virtually no selection occurs during the college course, so far as test intelligence is concerned. To be sure, there is some selection in terms of scholastic record, there being a superiority of 5.46 points in the freshman scholastic rating of those who persisted over those who did not become seniors. The critical ratio of this difference is 2.73.

9. The I.Q.'s of the Oberlin freshmen range from 92 to 169. 99 per cent of the I.Q.'s are over 100. Variability is represented by a sigma of 10.4, as compared with 16.4 for the Terman-Merrill standardization group.

10. Although the correlation between I.Q. and college scholarship is .40, the range of scholastic performance is remarkably similar at the different test levels between 101 and 140.

11. Four students with I.Q.'s between 91 and 100 became seniors, but their records were not brilliant.

12. Although the retentivity of the college was not materially greater for those at the extremely high end of the distribution than for those at the lower end, 80 per cent of those with I.Q.'s above 126, as compared with 64 per cent of those with I.Q.'s below 116, who had been in college long enough, became seniors.

Conclusions

Two facts of general significance emerge from the computations: First, the figures indicate that, although it is to be expected that students with higher intelligence test scores will make the better college records, it is nevertheless possible for the average of the group with I.Q.'s as low as 101-105 to do acceptable work at Oberlin. There are indeed exceptional students who, in spite of the handicap of an intelligence quotient as low as 92, obtain the A.B. degree. Second, test scores show a consistently closer correlation with college scholarship than with high school records. Interpretation of these facts would seem to point to the significance of adequate motiva-

tion. Possessed of determination, drive, and directionality, the student whose intellectual ability barely equals the average of the general population can "make the grade", if equipped with a good secondary school preparation. Selection of the student body at Oberlin is made primarily on the basis of the high school record. Students with low I.Q.'s, who rank in the lower half of their high school class, are not admitted. The higher validity figures obtained when college scholarship is used as the criterion also emphasize the factor of motivation. Oberlin students, at any rate, apparently work more nearly up to their potential capacity, so far as this is measured by the intelligence tests, while in college than in secondary school.

REFERENCES

1. Adams, F. J. "College Degrees and Elementary-School Intelligence Quotients", *Journal of Educational Psychology*, XXXI, (1940), 360-368.
2. Byrns, R. and Henmon, V. A. C. "Long Range Prediction of College Achievement", *School & Society*, XLI, (1935), 877-880.
3. Hartson, L. D. "Further Validation of the Rating Scales Used with Candidates for Admission to Oberlin College", *School & Society*, XLVI, (1937), 155-160.
4. Terman, L. M. and Merrill, M. A. *Measuring Intelligence*. Boston: Houghton Mifflin Company, 1937.

THE THURSTONE PRIMARY MENTAL ABILITIES TESTS AND COLLEGE MARKS

MARY LOU ELLISON

and

HAROLD A. EDGERTON

Ohio State University

THE PRESENT STUDY of Thurstone's Primary Mental Abilities Tests has been made in order to implement the assumption that the scores of the several factors might be useful in academic counseling. Four questions form the basis for the investigation.

1. What relationships are there between the factor scores and academic grades?

2. What relationships are there between the Ohio State University Psychological Test score and the factor scores?

3. How well can academic grades be predicted on the basis of the primary factor scores?

4. Are the factor scores related to grades in specific college subjects?

Thurstone's development of his Primary Mental Abilities Tests was for the purpose of appraising seven primary factors of mind¹ His isolation of these factors and the development of the final test battery is described in the monograph "Primary Mental Abilities."² Thurstone briefly describes the factors on his individual record sheet for the tests as follows:

"Factor P. The tests that call for this ability require the quick perception of detail in either visual or verbal material. This seems

¹L. L. Thurstone, *Manual of Instructions for Administering Tests for Primary Mental Abilities*, p. 2.

²L. L. Thurstone, "Primary Mental Abilities," *Psychometric Monographs* Chicago: The University of Chicago Press, 1 (1938).

to be a perceptual ability which enables some people to excel in finding detail which is significant to them or detail which they are seeking. It is probably one of the factors that is involved in what has been called 'quick intelligence.' Scanning a page to find quickly some small but significant detail and classifying familiar objects quickly are examples of this factor.

"Factor N This is one of the clearest factors that has been isolated. It consists of facility with simple numerical work and is best represented in the tests of rapid calculation. It is of secondary importance in arithmetical reasoning and in deciphering numerical code, tasks which call for factors in addition to facility with numbers as such. It is not yet known whether this factor can be exemplified in non-numerical tasks.

"Factor V. This is a verbal factor which is manifested in tests that involve the interpretation of language. It is not restricted to mere fluency with words. It reflects an ability to deal readily and quickly with verbal material. Those who excel in this factor are probably verbally-minded in their thinking and problem-solving.

"Factor S. This is an ability that is present in those tests which require the subject to think visually of geometrical forms and of objects in space. While none of these factors can be described in detail yet, it seems reasonable to expect that those who have a high rating on ability S should be able to do well in those studies and in those occupations that require visualizing or thinking about things in visual form. Many people think about a problem visually even when the nature of the problem does not immediately suggest any necessary visual character.

"Factor M. The nature of this factor was identified by the fact that all of the tests which require it are tests of memorizing. The appearance of such a factor seems to give justification for the belief that a good memory is an ability independent of other mental powers. It is not yet known, however, whether the ability to memorize is the same as the ability to recall experiences which we do not intend to retain for future recall. The present factor **M** can be tentatively named the ability to memorize.

"Factor I. The tests which require this factor demand that the subject discover some rule or principle in the material of the test. The factor does not seem to be restricted to material which is primarily numerical, primarily visual, or primarily verbal, types which were all represented in the tests for this factor. The ability

THURSTONE PRIMARY MENTAL ABILITIES TESTS

to discover a rule or principle in the solution of a problem is usually called induction. People differ markedly in the kind of resourcefulness that is involved in inductive thinking, and the hypothesis that the factor I is associated with this kind of ability seems plausible. It is not known whether this factor is associated with inventiveness and initiative.

"Factor D. The deductive factor is still only tentatively identified. It is a factor which is present in syllogistic reasoning and also in some other tests. It is one of several factors that may be involved in restrictive thinking. In a general description, the factor seems to represent facility in formal reasoning."

In the present study, Thurstone's Primary Mental Abilities Tests, Experimental Edition were used

The subjects consisted of a group of 49 students in the College of Arts and Sciences, Ohio State University. Most of those who took the test were students in the Exploratory Program of the College of Arts and Sciences.

The students tested do not constitute a random sample of students of the Exploratory Program, nor of the College of Arts and Sciences, nor of freshmen generally. This fact must be taken into consideration in the interpretation of the results of the study. No one was required to take the test. Of the forty-nine subjects, forty-one were freshmen, six were sophomores, and two were juniors. In the group, 39 per cent ranked in the 90th percentile or above in intelligence (Ohio State University Psychological Test), and 54 per cent were included in the 80th percentile or above. The mean Point Hour Ratio³ was 2.40.

In addition to the scores for the seven factors, and the separate scores on the sixteen individual tests from which the factor scores are derived, other data from the college records were used. Intelligence test percentiles were based on scores received in the Ohio State University Psychological Examination, given to all students at the time of entrance

³The Point Hour Ratio is the total points divided by the hours attempted. For each hour of grade A, four points are given; for each hour of B, three points; C, two points; D, one point; and E (failure), zero points.

to the University. The Point Hour Ratio for each student was obtained. Grades received in English, sciences, foreign languages, and psychology were recorded, since it was thought that each of these groups might be related differentially to the factor scores. In the group of forty-nine students, English grades were available for twenty-seven, science grades for thirty, foreign language grades for twenty-seven, and psychology grades for twenty-five.

1. What relationships are there between the factor scores and Point Hour Ratio?

In Table 1, the correlations between Point Hour Ratio and the various factors are shown. The correlation between Factor V and Point Hour Ratio is the highest (0.44). Factor M ranks second in its correlation with P. H. R., the correlation being 0.31. The other five factors have correlations with P. H. R. ranging from -0.24 to 0.19 . One might speculate on the meaning of the negative correlations, but on the basis of such a sample it might be unfortunate.

It is likely that in a really random sample of University students or of University freshmen such correlations would be zero or positive.

The multiple correlation between P. H. R. and the weighted scores of the seven factors is 0.640. When the Ohio State University Intelligence Test score is included as a variable with the seven factors, the multiple correlation is 0.648. Such a correlation suggests that there may be some justification for the use of the Primary Mental Abilities Tests for the prediction of academic success in college.

2. What relationships are there between the Ohio State University Psychological Examination scores and the factor scores?

As in the case with Point Hour Ratio, Factor V shows the highest correlation with intelligence (0.52). This is perhaps due to the fact that the expression of intelligence is largely verbal in character in present tests. The Same-Opposite Test, a component of Factor V, shows the highest correlation of the several sub-tests with intelligence test

THURSTONE PRIMARY MENTAL ABILITIES TESTS

TABLE 1

COMPARISON OF CORRELATIONS OF FACTORS AND INDIVIDUAL TESTS
WITH INTELLIGENCE TEST SCORES AND POINT HOUR RATIO
(N = 49)

	Intelligence Test		Point Hour Ratio	
	Composite Factor	Individual Test	Composite Factor	Individual Test
Factor P	0.06		-0.24	
Identical Forms		-0.05		-0.21
Verbal Enumeration		0.16		-0.27
Factor N	-0.02		0.17	
Addition		-0.07		-0.07
Multiplication		0.01		0.29
Factor V	0.52		0.44	
Completion		0.41		0.33
Same-Opposite		0.55		0.37
Factor S	-0.11		-0.21	
Figures		-0.12		-0.40
Cards		-0.07		-0.01
Factor M	0.28		0.31	
Initials		0.34		0.32
Word-Number		0.09		0.17
Factor I	0.11		-0.13	
Letter Grouping		0.24		-0.18
Marks		0.04		-0.32
Number Patterns		0.07		-0.23
Factor D	0.10		0.19	
Arithmetic		0.09		0.10
Number Series		0.36		0.35
Mechanical Movement		-0.13		0.04

scores. A somewhat similar test is found in the Ohio State University Psychological Test.

The correlation of Factor M with intelligence is 0.28. The Initials Test correlated 0.32 with intelligence, while the other component of Factor M, the Word-Number Test, correlated very low (0.09).

The correlation of factors P, I, and D with intelligence are positive, but are very low. Among the components of Factor D, the Arithmetic Test has a low correlation with intelligence (0.09), the Number Series Test has one of the highest correlations in the battery with intelligence, and the Mechanical Movements Test shows a negative correlation

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 2
INTERCORRELATIONS BETWEEN POINT HOUR RATIO, PSYCHOLOGICAL
TEST, AND THE SEVEN FACTORS
(N = 49)

	(Intelligence) O.S.U. Psych. Test	P	N	V	S	M	I	D
P. H. R.	0.20	-0.24	0.17	0.44	-0.21	0.31	-0.13	0.19
O. S. U. Psych. Test	0.20	0.06	-0.02	0.52	-0.11	0.28	0.11	0.10
P	-0.24	0.06	0.34	0.27	0.54	0.14	0.47	0.10
N	0.17	-0.02	0.34	0.43	0.18	0.28	0.38	0.42
V	0.44	0.52	0.27	0.43	0.12	0.33	0.36	0.34
S	-0.21	-0.11	0.54	0.18	0.12	0.08	0.34	0.19
M	0.31	0.28	0.14	0.28	0.33	0.08	0.13	0.14
I	-0.13	0.11	0.47	0.38	0.36	0.34	0.13	0.17
D	0.19	0.10	0.10	0.42	0.34	0.19	0.14	0.17
Mean	2.40	71.6	139.9	107.1	76.9	108.7	15.1	30.7
Standard Deviation	0.59	20.3	25.1	36.4	23.8	36.4	7.6	7.8
								2.4

with intelligence. Such correlations might raise a question regarding the functional unity of the factors

3. *How well can Point Hour Ratio be predicted on the basis of the primary factor scores?*

It would be desirable to be able to predict the probable P. H. R. of a student from the scores made on the seven factors. The chart below shows that in both situations, the highest beta weight is that for Factor V.

TABLE 3
BETA AND b REGRESSION COEFFICIENTS
For the Scores When the OSU Psychological Test is Included
and When It Is Omitted From the Test Battery

	OSU Intelligence Included		OSU Intelligence Omitted	
	Beta Coefficient	b Coefficient	Beta Coefficient	b Coefficient
Factor P	-.279	-.007	-.291	-.007
Factor N	.034	.001	.090	.001
Factor V	.568	.014	.487	.012
Factor S	-.113	-.002	-.089	-.001
Factor M	.216	.017	.191	.015
Factor I	-.196	-.015	-.201	-.015
Factor D	.046	.045	.040	.039
OSU Psychological Examination	-.136	-.004		

THURSTONE PRIMARY MENTAL ABILITIES TESTS

The correlation with P. H. R. is increased slightly when the intelligence test rating is used, the correlation between P. H. R. and the variables being raised from 0.640 to 0.648. In a random sample of freshmen this difference would probably be greater.

4. *Are the factor scores related to grades in specific college subjects?*

The correlations of course grades with Point Hour Ratio, intelligence test scores, and the seven factors are found in Table 4. The grades taken into consideration in this study are those in English, science, foreign languages, and psychol-

TABLE 4

THE CORRELATION OF SUBJECT MATTER GRADES WITH POINT HOUR RATIO, INTELLIGENCE, AND THE SEVEN FACTORS

	English Grade	Science Grade	Foreign Language Grade	Psychology Grade
P. H. R.	0.72	0.85	0.77	0.58
Intelligence	0.42	0.42	0.54	0.02
Factor P	0.10	-0.12	0.27	0.10
Factor N	0.34	0.03	0.45	0.37
Factor V	0.75	0.68	0.44	0.59
Factor S	0.44	0.23	0.56	-0.08
Factor M	0.42	0.18	0.45	0.23
Factor I	0.24	0.05	0.78	0.06
Factor D	0.44	0.23	0.43	0.63
Number of Cases	27	30	27	25

ogy. The results must be taken as suggestive and not as facts from which broad generalizations may be drawn.

In all four cases, there are high correlations between P. H. R. and grades. This is to be expected, since these grades are components of the Point Hour Ratio.

English grades correlate highest with Factor V (0.75). Factors S, M, and D also show correlations above 0.40 with English grades.

The only factor showing a correlation above 0.40 with science grades is factor V.

All the factors are apparently important in determining foreign language grades, since all factors except Factor P correlate above 0.40 with foreign language. The most significant correlation with foreign language is Factor I (0.78). This correlation is higher than one would expect, but it may be due to the fact that an inductive method is used at Ohio State University in teaching the beginning language courses.

The highest correlation among the factors with psychology grades is with Factor D (0.63). Factor V is also high (0.59). These were the only two factors correlating above 0.40 with psychology.

Factor V correlates above 0.40 with grades in each of the four subject fields considered, the highest being with English grades. The correlations between Factor I and the school subjects are low with the exception of foreign language grade (0.78). Factor P shows very low correlations with all four school grades. There is little differentiation between the correlations of the school grades and Factor N, the only correlation higher than 0.40 being with foreign language grade. Factors S and N both have correlations over 0.40 with English and foreign language grades, and Factor D has a significant correlation with English, foreign language, and psychology grades.

Such observations as reported here suggest that, with more experience, the Thurstone Primary Abilities Test will become a useful instrument in the academic counseling program of colleges. It will be necessary to secure more data in regard to the relationships of test scores and course grades from a random sample of freshmen. Also, it will be important to have some knowledge of methods of instruction in the several courses so as to judge whether the relationship observed is a function of the abilities of the student and the subject matter being studied, or of the methods of instruction.

A SHORT CUT IN THE ESTIMATION OF SPLIT-HALVES COEFFICIENTS

CHARLES I. MOSIER

Social Security Board

FOR SEVERAL YEARS the writer has availed himself of a short cut in the computation of reliability coefficients by the split-halves technique. The method has probably been developed independently by a number of other investigators, but it has not, to the writer's knowledge, appeared in print in connection with this specific problem, and there may be some workers to whom it may prove useful.

With the development of the Kuder-Richardson method for the determination of reliability, the split-halves technique should probably disappear from the scene. However, as a number of investigators have found, it provides a fairly close approximation to the Kuder-Richardson value, and since it does not require an item-analysis it will probably continue in use. In any event, the purpose of this note is not the justification of the technique, but the presentation of a short cut. If split-halves coefficients are to be computed, they may as well be computed efficiently.

In brief, the short cut involves the use of the complete dependence of the "even" scores on the "total" and the "odd" scores. We may suppose that "total" scores have already been obtained in connection with the original purpose of the test. Because of this algebraic dependency, then, it remains only to rescore the papers for the "odd" scores in order to know the even scores, since,

$$E_i = T_i - O_i \quad (1)$$

Furthermore, equation (1) need not be applied to each case separately. Not only may we dispense with the necessity of

rescoring the papers to get the "even" score for each individual; we may go farther, and dispense with the necessity of obtaining the individual "even" scores at all.

By defining the value desired, namely, the correlation between O and E , and substituting the value of E expressed as a function of O and T from equation 1, we obtain the result that

$$r_{OE} = \frac{r_{OT} \sigma_T - \sigma_O}{\sqrt{\sigma_T^2 + \sigma_O^2 - 2r_{OT} \sigma_O \sigma_T}} \quad (2)$$

This expression calls for only the "odd" and "total" scores, from which are obtained their respective standard deviations and the correlation between them. The value obtained from the odd-even correlation can then be used in the Spearman-Brown formula to give the estimated reliability. This value is identical with that which would be obtained if each test paper were independently scored for the "even" score, and the odd-even correlation coefficient computed from the resulting data. (This can be seen from the derivation.)

The expression in equation (2) is readily recognized as a special case of the more general formula for a correlation between a part and the whole exclusive of the part. O , E , and T in equation (2) are by no means limited to "odd," "even" and "total" scores, but apply to any set of variables for which equation (1) is true.

MEASUREMENT ABSTRACTS*

Adams, C. R. "A New Measure of Personality." *Journal of Applied Psychology*, XXV (1941), 141-151.

A new instrument for measuring personality traits, The Personal Audit, is described. It was intended to be relatively free from highly personal items since it was felt that such a test would be more useful in non-clinical situations. The Personal Audit is believed, on the basis of low intercorrelations between sub-tests, to measure 9 relatively independent personality traits. Coefficients of reliability (corrected split-half) range from $+.90$ to $+.96$. Items have been validated by the criterion of internal consistency using a modified version of the Sletto technique *W. A. Varvel*.

Baxter, B. and Paterson, D. G. "A New Ratio for Clinical Counselors." *Journal of Consulting Psychology*, V (1941), 123-126.

The magnitude of the $S.E._M$, which clinical counselors employ in interpreting test scores, varies in significance with the variability ($S.D.$) of the norm group. It is useful, therefore, to relate them in a ratio as an aid in interpreting scores. The following formula, in which r is the reliability coefficient, provides a simple way of expressing the magnitude of $S.E._M$ as a percentage of $S.D.$.

$$\frac{S.E._M}{S.D.} = \frac{S.D. \sqrt{1-r}}{S.D.} = \sqrt{1-r}$$

Application of this ratio to a list of 49 tests shows that $\frac{S.E._M}{S.D.}$ ranges from as low as .10 to as high as .55. In general, achievement tests show the lowest ratio (highest accuracy) with an average of .20, followed in order by scholastic aptitude tests (averaging .30), reading tests (.32), special

*Edited by Forrest A. Kingsbury

aptitude tests (.33), and personality tests (.27 to .55, average .40). *F. A. Kingsbury.*

Bennett, G. K. and Raskow, S. "Extension of the Norms of the Columbia Vocabulary Test" *Journal of Applied Psychology*, XXV (1941), 48-51.

Constructed and standardized for grades 3-8, this test showed a mean score of 54 and standard deviation of 15 for the latter half of the eighth grade. Extension of norms seems justified when 1212 superior recent high school graduates obtained a mean score of 74 with standard deviation of 12. When the test was administered to 5101 high school students, mean scores increased and the standard deviation decreased from grade 9A through grade 12B both among commercial and general-course students. Decreasing standard deviation probably indicates increasing homogeneity of vocabulary in later school years. With grade constant, mean scores decrease with age. *J. E. P. Libby.*

Blum, Milton L. and Candee, Beatrice. "The Selection of Department Store Packers and Wrappers with the Aid of Certain Psychological Tests: Study II." *Journal of Applied Psychology*, XXV (1941), 291-299.

This study is a check on conflicting results in previous attempts to determine the value of finger dexterity tests in predicting successful wrappers or packers. Tests used were the O'Connor Finger Dexterity, Zeigler Placing, Otis Self-Administering, and Minnesota Clerical. Test performance was checked against production records and foreman's ratings. Results indicate no relation between finger dexterity and production for either packers or wrappers. In the experienced group the Minnesota Clerical shows positive correlation for both groups. It is concluded that clerical speed and accuracy have a much higher relation to production than has finger dexterity. *D. A. Peterson.*

Brown, A. W. and Blakey, R. "A Preliminary Report on the Development and Standardization of a Non-Verbal Test at the High-School Level." *Journal of Educational Psychology*, XXXII (1941), 113-123.

A series of 11 non-verbal subtests constructed on the concepts of primary mental abilities has been standardized on a group of 286 suburban high school students. Eight of these subtests, two of perceptual speed, two of spatial relations, and four of abstract reasoning, constitute the final test, which may be given in forty minutes. The "Non-Verbal Reasoning Test" correlates with school grades .47 and with Otis I.Q. .59; Otis I.Q. correlates with school grades .60. Higher correlation with grades was not expected since the latter involve other abilities in addition to those in the non-verbal test. Reliability of the test is .97. Tentative norms are given, including derived scores intended to take the place of I.Q.'s at this level; standardization on a much larger sample is being undertaken. *J. E. P. Libby*.

Brown, A. W. and Cotton, C. B. "A Study of the Intelligence of Italian and Polish School Children from Deteriorated and Non-Deteriorated Areas of Chicago as Measured by the Chicago Non-Verbal Examination." *Child Development*, XII (1941), 21-30.

1262 Italian and Polish school children in a deteriorated and a non-deteriorated area were tested with a non-language group test battery. The children were in the fourth grade or above and from 10 to 14 years of age. The authors stress the influence of socio-economic, cultural, and educational factors upon test scores. They found (1) a regular decrease in mean test performance from age 10 to age 14 for both sexes and both nationality groups but not so great as that previously reported for verbal tests; (2) sexual differences favoring the boys, particularly in the case of Italian children; and (3) contradictory indications relating to socio-economic community level (no significant differences between areas for Italian boys; significant differences favoring the deteriorated area for

Italian girls; a tendency for Polish children in the non-deteriorated area to make better scores). *W. A. Varvel*.

Brush, Edward N. "Mechanical Ability as a Factor in Engineering Aptitude." *Journal of Applied Psychology*, XXV (1941), 300-312.

This study was intended to explore the possibilities of available tests of mechanical ability and aptitude as indicators of aptitude for engineering. The report is prefaced with a survey of the relevant literature. The subjects were two groups of students in the College of Technology at the University of Maine, one group of 104 members, the other group of about 130 members. The criterion was scholastic rank in courses of an engineering nature. The tests used were: Minnesota Paper Form Board, Minnesota Assembly Test, Minnesota Spatial Relations Test, O'Connor Worksample No. 1, O'Connor Worksample No. 5, O'Connor Worksample No. 72, Cox Mechanical Explanation and Completion Test, Cox Mechanical Models Test, and MacQuarrie Test for Mechanical Ability. In addition data on intelligence tests, algebra, chemistry, plane geometry, and physics tests were also available.

The conclusions reached are summarized as follows: "The tests of useful predictive power were the Cox Tests of Mechanical Aptitude and Minnesota Paper Form Board . . . Batteries of mechanical ability tests yield correlations with the criterion of about .40; batteries in which an intelligence test is combined with one or two tests of mechanical ability yield correlations of about .50 . . . several batteries of mechanical ability tests predict engineering scholarship at least as well as the intelligence tests, while the achievement tests, singly and in combination, predict success in engineering studies somewhat better than do the tests of mechanical ability . . . total engineering record is more highly correlated with first semester and first year grades than with any test or combination of tests." *J. E. Karlin*.

Burt, H. E. and others. "Market Problems and Market Research." *Journal of Consulting Psychology*, V (1941), No. 4, 145-193.

This entire number is devoted to eight papers on market research, not separately abstracted here because of space limitations. The authors and titles are as follows: "Current Trends in Marketing Research" (H. E. Burt); "Proving Ground on Public Opinion" (H. G. Weaver); "Problems of Sampling in Market Research" (Frank Stanton); "Characteristics of the Question as Determinants of Dependability" (J. G. Jenkins); "Evaluating the Effectiveness of Advertising by Direct Interviews" (P. F. Lazarsfeld); "Effects of Repeated Interviewing on the Respondent's Answers" (F. D. Ruch); "The Museum Technique Applied to Market Research" (G. K. Bennett); and "The Role of Psychological Interpretation in Market Research" (A. W. Kornhauser). *F. A. Kingsbury*.

Casanova, T. "Analysis of the Effect upon the Reliability Coefficient of Changes in Variables Involved in the Estimation of Test Reliability." *Journal of Experimental Education*, IX (1941), 219-228.

The following topics are discussed and various formulae developed in detail: (1) the variance of the halves in the split-half method of estimating reliability; (2) the correction for guessing with specific reference to the reliability of rights and wrongs, the variance of rights and wrongs, the correlation of rights with wrongs, the variance of the number of items attempted, and the number of possible choices; (3) the effect of calling all negative scores zero; (4) the variance of the items. In the latter case, a formula for estimating the reliability of a test in terms of the item variances is presented which is felt to be more convenient than the Kuder-Richardson formulae. *W. R. Varvel*.

Cattell, Raymond B., Feingold, S. Norman, and Sarason, Seymour B. "A Culture-Free Intelligence Test: II. Evaluation

of Cultural Influence on Test Performance." *Journal of Educational Psychology*, XXXII (1941), 81-100.

A culture-free intelligence test described in an earlier paper was administered together with the Binet (Terman-Merrill), A.C.E. (arithmetical sections), and the Arthur Performance Tests, to four comparable groups; these were given special training, one group in each class of information or skill demanded by the tests. Retest analyses showed the Arthur least influenced by training in its own culture medium, the Culture-Free Test next, Binet next, and A.C.E. most influenced. The above tests and the Ferguson formboards were administered to a group of adult immigrants, resident in this country about one year, and a control native group. The Ferguson was very close to the Arthur, others followed in the order noted earlier, when the groups were retested after 77 days during which the immigrants gained noticeably in Americanization. Reliability of the Culture-Free Test compares favorably with those of the others. Adequate validity is indicated by the Culture-Free Test's high loading in the general factor brought out by tetrads, and by its high mean correlation with the pool of tests. Since life experience probably brings factors in the Culture-Free Test to saturation in widely different cultures, its proper application appears broader than that of preceding tests. *J. E. P. Libby.*

Driver, Randolph S. "The Validity and Reliability of Ratings." *Personnel*, XVII (1941), 185-191.

Rating is of value in industry only when its limitations as a scientific instrument are fully appreciated. The various current methods of obtaining measures of validity and reliability are discussed and their values and limitations considered. In order for a rating to be acceptable, it must be proven valid and reliable. Although difficult to accomplish, ratings are not useless, but great caution must be observed in their interpretation. *Virginia Brown.*

Dudycha, George J. "A Suggestion for Interviewing for Dependability Based on Student Behavior." *The Journal of Applied Psychology*, XXV (1941), 227-231.

College students were divided into groups of extreme earliness and lateness, of dependability and undependability, on the basis of observation of their behavior in life situations. Ten questions on punctuality and persistence and six on dependability, when presented to these contrasting groups, elicited responses indicating significant group differences. Since these questions appear to be diagnostic in student behavior, it is suggested that they be tested for usefulness in employment situations for discovering those applicants likely to prove dependable. *Virginia Brown.*

Dulsky, S. G. "Vocational Counseling. I. By Use of Tests; II. By Interview." *Personnel Journal*, XX (1941), 16-28.

The author briefly and critically examines various types of standardized tests available to the vocational counselor. He concludes that aptitude tests are of no value and personality tests of very limited value. Interest inventories, if used properly, may be helpful. Tests of intelligence and educational achievement are approved as being of the most value. He advocates greater emphasis on the vocational interview as a means of diagnosing personality and motivation and of identifying and evaluating interests. Self-guidance from the study of test scores and profiles is impossible. Vocational counseling is an individual process, requiring "skilled psychologists" rather than "mental testers." The vocational counselor should confine himself to descriptive rather than quantitative reports of test and interview results and should only rarely go beyond general recommendations to his clients. *W. A. Varvel.*

Ebert, Elizabeth H. "A Comparison of the Original and Revised Stanford-Binet Scales." *The Journal of Psychology*, XI (1941), 47-61.

1434 records of 315 children five to ten years of age were studied for information as to the comparability of I.Q.'s from the original and revised Stanford-Binet Scales. An increasing

discrepancy between I.Q. values on the two scales was found at ages 7, 8, and 9. The new revision tends to give lower I.Q.'s for levels below 100 and higher I.Q.'s for levels above 100. The duller children gain slightly more in I.Q. than the brighter ones although both groups show increases. With the old revision, the duller individuals gain but the brighter ones lose. Although the average I.Q. of the 1916 revision was more constant, individuals maintained their relative positions better in the new revision. *Virginia Brown.*

Eysenck, H. J. "Type-Factors in Aesthetic Judgments."
British Journal of Psychology, XXXI (1941), 262-270.

It has been found previously that the analysis of the inter-correlations between the rankings of pictures by a number of subjects yields mainly one general factor with no other significant factor. On this occasion the attempt is made to bring out the influence of any such secondary factor, even, if need be, at the expense of the "T" or general factor. Five series of pictures, each consisting of thirty to fifty items, were judged in order of goodness by fifteen subjects. The subjects were artists, university students, bank clerks, typists, and teachers, eight women and seven men, with age range from 20 to 70. The table of correlations for each of the five series was factored and two significant factors extracted in all cases except one. One factor was the "T" factor previously identified; the other factor, called the "K" factor, seemed to divide the population into two different "types," one preferring the modern, and the other the older style of painting. This factor, identified provisionally with "brightness," correlated with extroversion, radicalism, youth, and possibly with preference for color. The color-form test also appeared to be correlated with extroversion. Results are definite enough to suggest that further research into the relation between temperament and aesthetic preferences will not only extend knowledge of the "type" factors in aesthetic judgments, but also increase understanding of temperamental "types." *J. E. Karlin.*

Ferguson, L. W. "A Study of the Likert Technique of Attitude Scale Construction." *Journal of Social Psychology*, XIII (1941), 51-57.

The suggestion is here examined that Likert's method of constructing and scoring attitude scales gives results as valid as those of the method outlined by Thurstone and Chave with much less labor. Items constructed by the former method (Minnesota Scale for the Survey of Opinions) were rescaled by the latter; standard deviations of the distribution of scale values indicate that such items are adequately scaled by this method. The scale values obtained indicate that Likert's technique does not obviate the need of a judging group. With one exception, the scales cannot be scored by the Thurstone method. Scores obtained by the two methods for the exceptional scale show a correlation of .70, confirming the conclusion. *J. E. P. Libby.*

Greene, E. B. *Measurements of Human Behavior*. New York: The Odyssey Press. pp.777. 1941.

This volume of 24 chapters is divided into three parts: Part I, "Basic Considerations" (discussing introductory concepts, varieties of appraisals, score-interpretation, measures of relationship, types of instruments, item construction and evaluation, factor analysis); Part II, "Instruments and Results" (tests of early childhood, of achievement, Binet-type and group intelligence scales, performance, mechanical and motor tests, measures of fine arts—design, literature, and music—tests of interests, attitudes, adjustment); Part III, "Persistent Problems" (effects of practice on scores, measures of growth and senescence, absolute scaling, evaluation of judgments, native differences). A 30-page bibliography, a combined glossary and subject-index, 121 tables, and 108 figures are features of the book. *F. A. Kingsbury.*

Guilford, J. P. "A Note on Dubois's Method of Deriving an Achievement Ratio for Students." *Journal of Educational Psychology*, XXXII (1941), 220-222.

Dubois's achievement ratio is that of the student's actual

average mark to that mark corresponding to the standard score obtained on a psychological test; these ratios in general are low for students with high test scores and high for those with low scores. This finding follows from the assumption of a correlation of 1.00 between test scores and marks, while Dubois gives the correlation as .442. A student may be expected to deviate from the mean mark by only .442 as much as his standard score indicates. It is suggested that Dubois's conclusion might be reversed if the regression line r equals .442 be taken as base. Computation of a special but meaningful case confirms this suggestion. *J. E. P. Libby.*

Hay, Edward N. "Tests in Industry." *Personnel Journal*, XX (1941), 3-15.

This is a discussion of the opportunities for psychologists in industry. The use of intelligence tests is coming more and more to act as a check on employer's judgment, which is customarily biased in favor of the qualities of aggressiveness and good personality. Such tests indicate the level at which the employee is able to work most efficiently and his potentialities for further promotion. It is particularly important to obtain psychological information about an employee at the time of his entry into a firm since the work he does then determines to a large extent his opportunities for advancement. A beginning job may require an I.Q. of about 100 but higher positions require higher I.Q.'s so that an employee progressing reasonably well in the initial job may become unfit when advanced to the more complex positions. It becomes advisable to judge prospective employees not on the basis of the intelligence required for their first positions but for the positions to which they should be able to rise. With the use of objective tests, information becomes generally available for an entire firm so that transfers and promotions from one department to another can be advised with a minimum of further consultation, since the qualities required in other work are known and the abilities of the employee are likewise known at the time of first testing. Apart from the question of job maladjustment

there is a further fruitful field for the industrial psychologist in the problem of a better supervisor-employee relationship. An illustrative study of these methods at work accompanies the discussion. *J. E. Karlin.*

Johnson, Donald M. and Reynolds, Floyd. "A Factor Analysis of Verbal Ability." *Psychological Record*, IV (1941), 183-195.

The literature on problem solving among animal and human subjects suggests that there may be two fundamental processes involved: "F," the flow of various acts or responses; and "S," the selection of these responses according to the requirements of the problem. This study tested the hypothesis that individual differences in these two processes is a major determinant for scores on problem-solving tests. This investigation was limited to verbal problems. There were ten tests involving the supplying of verbal responses; the tests varied in restriction of choice of responses from complete freedom to supply any word to restriction to the supplying of only certain words according to a rigid criterion. The subjects were 113 summer-school students at Fort Hays Kansas State College. A centroid analysis of the table of corrected correlation coefficients yielded two factors. The tests fell within a positive manifold, after rotation, indicating two definite factors reasonably identified as the "F" and "S" postulated in the hypothesis. It appears that these two factors are probably closely related to, if not identical with, Thurstone's "W" and "V" factors. It is concluded that the two processes or functions mentioned account to a large extent for the variance in verbal problem-solving tests. These findings are further discussed with reference to tests of vocabulary, intelligence, and reading. *J. E. Karlin.*

Kornhauser, A. W. and Schultz, R. S. (et al). "Research on Selection of Salesman" (and other papers). *Journal of Applied Psychology*, XXV (1941), No. 1, 1-47.

Five papers read at the Section of Industrial and Business Psychology of the American Association for Applied Psychol-

ogy in 1940, together with an introductory article, are presented in this number, but are not separately abstracted because of space limitations. In addition to the introductory article (title as above), the authors and titles of the papers are as follows: "Selection of Casualty and Life Insurance Agents" (M. A. Bills); "Recent Research in the Selection of Life Insurance Salesmen" (A. K. Kurtz); "A Report of Research on the Selection of Salesmen at the Tremco Manufacturing Company" (O. A. Ohmann); "Procedures for the Selection of Salesmen for a Detergent Company" (J. L. Otis); and "Selection Research in a Sales Organization" (T. M. Stokes). *P. A. Kingsbury.*

Lowell, Frances E. "A Study of the Variability of I.Q.'s in Retests." *Journal of Applied Psychology*, XXV (1941), 341-356.

The main purpose of this study was to seek corroboration of the results obtained in Cleveland Public Schools in recent years which seemed to show that the I.Q.'s of school children tended in certain instances to vary between test and retest. The data were composed of 1000 cases that had two tests only, 1000 cases that had three tests, and 1000 cases that had four tests. The Terman 1916 revision of the Binet was used in all tests. It was found that there are significant decrements in I.Q. both for groups and for chronological age. Furthermore, the I.Q. range, the chronological age at first test, and the interval elapsing between first and last tests may all be eliminated as causes for variation in I.Q. on retest. Nor does sex influence variations in I.Q. between first and last tests. On the average, four times as many cases on retest decrease in I.Q. as increase. In particular, those cases that increase 7 or more points on the first retest decrease 5 times as often as they increase on the second retest. The data on the first retest seem to indicate that the older the child is, the less chance there is that his second I.Q. will increase. *J. E. Karlin.*

McCloy, C. H. "The Factor Analysis as a Research Technique." *Research Quarterly*, XII (1941), 22-33.

MEASUREMENT ABSTRACTS

This paper presents an elementary discussion of some fundamental concepts and limitations of factor analysis. Particular reference is made to its possible uses in the field of health and physical education. Specific examples of precautions to be taken and the kind of studies to which this type of correlational analysis may be applied are given in terms of research in physical education. The method has been utilized in (1) studies of motor skills, (2) analysis of anthropometric data, (3) analysis of cardiovascular variables, and (4) studies of character and personality traits. A 17-item bibliography is included. *W. A. Varvel.*

Mosier, C. I. "A Psychometric Study of Meaning." *Journal of Social Psychology*, XIII (1941), 123-140

256 adjectives expressing judgmental relationships which could be placed along a favorable-neutral-unfavorable continuum were rated on an 11-point scale by college students in psychology. Some 140 ratings were obtained for each word "Two basic hypotheses . . . are confirmed: first, that the meaning of a word may be considered *as if* it consisted of two parts, one constant and representative of the usual meaning of the word, and one variable, representative of individual interpretation in usage and associated context and general usage, second, that the frequency with which any particular meaning is evoked is describable by the Gaussian Law" The presence of words with two discrete meanings, yielding bimodal frequency distributions of responses, was noted. The effect of adverbial modifiers on the meaning of an adjective was studied "A scale with a rational basis has been developed and values describing quantitatively the modal meaning and the ambiguity of more than 200 adjectives have been obtained." *W. A. Varvel.*

Oral Trade Tests—Group Leaders' Handbook The Personnel and Training Section in collaboration with the Local Office Operations Section and Chicago Occupational Research Center. Division of Placement and Unemployment

Compensation, 205 West Wacker Drive, Chicago, Illinois.
February, 1941. 18 pp.

This handbook is designed to instruct the group-leaders in interviewing in the construction and use of Oral Trade Tests. It is based on *Oral Trade Questions, Vol. I*, prepared by Occupational Analysis Section, United States Employment Service Division (not available for general use). There are three divisions. History of Oral Trade Questions; Preparation of Oral Trade Questions; and Application of Oral Trade Questions in Operating Offices. Specific examples of the application of oral trade questions are given. *D. A. Peterson.*

Osgood, C. E. and Stagner, Ross. "Analysis of a Prestige Frame of Reference by a Gradient Technique." *Journal of Applied Psychology*, XXV (1941), 275-290.

This study was designed to demonstrate a method for analyzing a frame of reference, and to investigate the particular determinants of the frame of reference known as occupational prestige. Subjects were required to judge a number of occupational stereotypes with respect to a psychological "gradient" continuum varying from the description "brains" on the one extreme to "brawn" on the other. In a second part of the test the judgments were made about persons rather than occupations. The subjects were 100 Dartmouth College men, students in introductory psychology, 50 of whom filled out the "job" form and 50 the "person" form. There were 15 names of occupations and each was accompanied by a set of ten characteristics. It was found that general rankings for prestige correlate on the average highly with median judgments on the gradient test, but that the reactions on the job forms were significantly different from the person forms. Prestige is imputed to occupations *per se* on the basis of such characteristics as hopefulness, being noticed, financial return, brains; prestige is imputed to men in specified jobs on the basis of brains, leadership, and self-assuredness. Since the conditions of the experiment are deemed to exclude the possibility of conscious verbalization of a prestige frame of reference, it is

concluded that the mere presentation of a set of occupational stereotypes for a series of judgments caused the spontaneous establishment of a prestige framework which then determined in a reliable manner judgments on the specific traits listed. The technique is practical and adaptable. *J. E. Karlin.*

Powell, N. J. "Check List for Use in Civil Service Objective Test Preparation." *Public Personnel Quarterly*, II (1941), 13-16.

The author has prepared a diagnostic check list designed to increase the probability of considering all the major bases for appraisal of the test being constructed. Guiding questions are listed under each of the construction problems for examining the individual item and the test as a whole. A dual criterion is suggested and instructions for use of check list are given. It is emphasized that while the degree of correlation between test score and job performance is important, it is not the only indicator of adequacy of examination. *D. A. Peterson.*

Powell, N. J. "Steps in Written Test Construction." *Public Personnel Quarterly*, II (1941), 73-76.

The process of constructing a written test is analyzed, assuming that examinations are made public (i.e., a test item cannot be used more than once). The following general problems are treated in outline form: 1. the determination of the abilities to be measured; 2. the determination of the test content which measures the desired abilities; 3. the allocation of emphasis; 4. the preparation of the test items; 5. the arrangement and editing of the test items; 6. the experimental tryout; 7. final test copy; and 8. general considerations with regard to test preparation integrity. *D. A. Peterson.*

Reyburn, H. A. and Taylor, J. G. "Some Factors in Intelligence." *British Journal of Psychology*, XXXI (1941), 249-261.

This study is intended to throw further light on the controversy regarding the unitary functioning of a general factor, *g*, in tests of intelligence. The material consisted of ten tests

purporting to measure some aspect of intelligence, the tests being formboards, repetition of digits, repetition of digits backwards, matching tests, absurdities, Porteus mazes, arithmetical reasoning, reasoning tests, vocabulary tests, and dissected sentences. The tests were given to 1497 South African children with ages ranging from 12 to 18. Five factors were extracted from centroid analysis of the inter-test correlations. The axes were then rotated orthogonally so as to preserve a positive manifold and, if possible, retain a general factor present in all the tests. It turns out, however, that no general factor is present. Three factors are immediate memory span (in digits forwards and digits backwards), verbal (in dissected sentences and vocabulary) and perceptual dexterity (in dissected sentences, matching, mazes); the two other factors are present in equal proportions in matching, arithmetic, and reasoning. Neither of these two is *g* as ordinarily operationally defined; one factor is the ability to find or make a significant pattern in a mass of irrelevant material, and the other factor is the ability of logical elimination. The suggestion is made that *g* in this battery is complex and that orthodox tests of *g* need to be constructed to preserve its functional unity. *J. E. Karlin.*

Roff, Merrill. "A Statistical Study of the Development of Intelligence Test Performance." *Journal of Psychology*, XI (1941), 371-386.

Using data available in the literature, correlations between test performance of children at a specific age and the gain in their performance one or more years later were estimated. The fact that the correlations showed no tendency to increase as the interval between test and retest increases indicates that the "Constancy of the I.Q." is due primarily to retention of earlier skills and knowledge rather than to correlations between earlier scores and later increments. On the assumption that the I.Q. variability is constant, the same procedures were used to find correlations which would result if scores and later increments were uncorrelated. No comparison of these values and empirical findings is made. *Lorraine Bouthilet.*

Schellhammer, Fred M. "The Intelligence Test in Teacher-Training Institutions." *School and Society*, LIII (1941), 319.

In a survey of 150 teacher-training institutions, 103 were found to use the intelligence test in student selection and evaluation of intelligence, 18 relying on it solely and 85 using it in conjunction with other techniques such as high school records and interview and faculty reports, no one combination finding universal favor. The majority of institutions considered the high school record as important as the intelligence test, and were supplementing both measures with subjective techniques. *Virginia Brown.*

Super, D. E. "A Comparison of the Diagnoses of a Graphologist with the Results of Psychological Tests." *Journal of Consulting Psychology*, V (1941), 127-133.

To check the claims of a woman "graphologist," 24 students submitted samples of their handwriting and obtained the graphologist's diagnoses. These were compared with the most appropriate of several test scores (Intelligence, Fryer & Sparling's Occupational Intelligence Norms, Strong Vocational Interest, and Bernreuter Personality Inventory). Use of chi-square and other methods showed no more than chance relationship between occupations recommended and those indicated as suitable for intelligence scores obtained; occupations rated as unsuitable by interest tests were recommended with more than chance frequency; personality traits were estimated by the graphologist with no more than chance agreement with test scores (on four traits), and worse than chance agreement (on two traits). *F. A. Kingsbury.*

Thomson, Godfrey. "Critical Notice of 'The Factors of the Mind' by Cyril Burt." *British Journal of Educational Psychology*, XI (1941), 45-51.

Thomson writes a brief review of Burt's most recent book (*The Factors of the Mind*, Univ. of London Press, 1940, xiv + 509). The major portion of the review considers Burt's

section on the distribution of temperamental types and the application of factor analysis to persons as well as to tests. Thomson does not agree that the philosophical approach to factor analysis is easier or more illuminating than the geometrical but he does express agreement with Burt's conclusions as to the metaphysical status of mental factors. W. A. Varvel.

Traxler, Arthur E. and others. "Psychological Tests and Their Uses." *Review of Educational Research*, XI (1941), 1-130.

This issue, consisting of eight papers not separately abstracted because of space limitations, is concerned with the construction, evaluation, and application of psychological tests. Individual articles are accompanied by extensive bibliographies. The following is a list of authors and titles:

- I "Brief Overview of the Period" (Arthur E. Traxler)
- II "Current Construction and Evaluation of Intelligence Tests" (Dewey B. Stuit)
- III "Applications of Intelligence Tests" (J. B. Stroud)
- IV "Measurement of Aptitudes in Specific Fields" (David Segel)
- V "Current Construction and Evaluation of Personality and Character Tests" (Arthur E. Traxler)
- VI "Projective Methods in the Study of Personality" (Percival M. Symonds)
- VII "Applications of Personality and Character Measurement" (John W. M. Rothney)
- VIII "Statistical Methods Related to Test Construction and Evaluation" (John C. Flanagan) D. A. Peterson.

Traxler, Arthur E. "Stability of Scores on the Primary Mental Abilities Tests." *School and Society*, LIII (1941), 255.

Test-retest correlations after one year ranging, with one exception, from .578 to .917 were found for the scores of 104 pupils in grades X-XII on Thurstone's Primary Mental Abilities Tests. The guidance value of the perceptual, memory, and inductive tests may be limited, for their correlations fell

below .80. These results should be checked with a larger and more representative group as the sampling and number of cases were not adequate in the present study. *Virginia Brown*

The Use of Tests in the Illinois State Employment Service

The Personnel and Training Section in collaboration with the Local Office Operations Section. Division of Placement and Unemployment Compensation, 205 West Wacker Drive, Chicago, Illinois. February, 1941. 11 pp.

This pamphlet is intended to assist interviewers in the use of test results as a supplementary tool in "making more objective the evaluations which must be made during the interview." The use of tests is related to other interviewer's tools (i.e., Job Descriptions, The Dictionary of Occupational Titles, Registration and Placements Aids). The article describes the types of tests, proficiency and aptitude tests, used in aiding the interviewer to evaluate work skills. Aptitude test batteries have been developed for three fields: selling, clerical work, and manual work. Three graphic illustrations of relation of scores on aptitude tests to job performance are given. *D. A. Peterson*

Viteles, M. S. "A Psychologist Looks at Job Evaluation." *Personnel Journal*, XVII (1941), 165-176.

The author recognizes the importance of job evaluation as a basic feature of the industrial relations program. In promoting the adjustment of workers, there is a need for a procedure designed to establish an equitable basis of compensation, to facilitate transfer and promotion, and to eliminate duplication of activities. The chief consideration of the paper is a critical examination of the various types of job evaluation programs in the light of psychological principles and experience. Ways are indicated in which improvements might be effected through the application of the techniques and principles of applied psychology. The present trial and error approach could be converted into a "rational, logical, and scientific system of analysis." *W. A. Varvel.*

Welch, Alfred C. "An Analytic System of Testing Competitive Advertising." *Journal of Applied Psychology*, XXV (1941), 176-189.

This study is intended to correct the usual copy-test procedure, which is defective in that it yields only a gross evaluation, by combining into a unified system a number of different tests which will provide the advertiser with clues to help him improve his advertising. An analytic system of testing competitive advertising was developed to provide a method of suggesting specific strong and weak aspects of an advertising campaign as well as to provide a gross evaluation of the effects of the campaign. The system is based upon four tests: A Brand Preference scale (described previously); a Brand Familiarity scale (controlled association or aided recall) in which the respondents were required to name five brands in response to each of two stimulus-words, cigarettes and fountain pens; a Theme Familiarity test (Link's method of triple associates) in which the respondent must identify the sponsor of a particular advertising theme; and a Theme Credence test (a belief test that does not require the respondent to report directly whether he believes an advertising claim). Tests of reliability and validity for the various scales indicate that the Brand Familiarity, Theme Familiarity, and Theme Credence Tests were useful supplements to the Brand Preference scale in analyzing the effects of advertising but that none of the three tests could be depended upon as a valid measure if used alone. Examples of the use of the analytic system are given. *J. E. Karlin.*

Wells, F. L. "Some Functions of Mental Measurements in the Young Superior Adult." *Journal of Consulting Psychology*, V (1941), 105-110.

A review of cases seen through the psychiatric division of the student health department in a large endowed university reveals about ten classes of adjustment problems. These are distinguished by different patterns of performance on the various standard examination techniques. Representative cases and usual methods of treatment for each class are described. *F. A. Kingsbury.*

